Neural and Behavioral Interactions in the Processing of Speech and Speaker Information

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)
im Fach Psychologie

Eingereicht am 11.12.2013 an der
Mathematisch-Naturwissenschaftlichen Fakultät II
der Humboldt-Universität zu Berlin
von **M.Sc., Jens Kreitewolf**

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät II
Prof. Dr. Elmar Kulke

Gutachter/Gutachterinnen:

1. Prof. Dr. Katharina von Kriegstein

2. Prof. Dr. Werner Sommer

3. Dr. Kai Alter

Tag der Verteidigung: 13.10.2014

Acknowledgments

First and foremost, I would like to thank my supervisor Katharina von Kriegstein for her constant support and confidence in my work. I am very thankful for her scientific guidance and all the things I had the chance to learn over the last four years. I would also like to thank all current and former members of the “Neural Mechanisms of Human Communication” group and everyone in the MPI who accompanied me during my PhD. In particular, I thank my former office mates and friends Sam Mathias, Katharina Schümberg and Katharina Stuke for supporting me during periods of frustration and reminding me that science can be a lot of fun. Furthermore, I thank Sam for proof-reading and always being there when I needed his help. I thank Etienne Gaudrain for introducing me to the magic of sound resynthesis. Finally, I would like to thank my parents and family for their unconditional love and Kamila Lübe for her emotional and selfless support. Without the contribution of the people mentioned here, this thesis would not have been possible. *Obrigado! Danke! Thanks! Merci!*

Table of Contents

1 Introduction

During natural conversation, we send rich acoustic signals that do not only determine the content of conversation but also provide a wealth of information about the person speaking, such as gender (e.g., Lass et al., 1976), approximate age (e.g., Shipp and Hollien, 1969), emotional state (e.g., Frick, 1985), and region of origin (e.g., Labov, 1972). Even in situations where only auditory information is available – during phone calls, for instance – normal hearing listeners are able to readily understand what is said as well as recognize who is speaking.

Although acoustic features specific to a speaker's voice are informative, they also impose large amounts of variability on the speech signal. Listeners are generally faced with the problem that the acoustic properties of the same speech message can be dramatically different depending on the way different speakers produce speech (Peterson and Barney, 1952), and that information relevant to both speech message and speaker identity is fundamentally intermingled in the same acoustic parameters. Nonetheless, listeners understand speech from a variety of different speakers with apparent ease (Abramson and Cooper, 1959; Peterson and Barney, 1952). The brain mechanisms that enable us to robustly understand speech under such variable conditions are still largely unknown.

Traditionally, the question of how we understand what is said has been studied separately from the question of how we recognize who is speaking (reviewed by Nygaard, 2005). Most of what we currently know about the neural basis of speech recognition comes from studies investigating speech from a single speaker (e.g., Hickok and Poeppel, 2007; Friederici, 2011; Scott and Johnsrude, 2003). Similarly, speaker-related attributes in the speech signal have been studied with little attention to linguistic processes (e.g., Belin et al., 2004; Van Lancker et al., 1985). Implicit in this empirical separation is the notion that speech information and speaker information are processed independently from one another. In recent years, however, evidence has accumulated that suggests an integrated neural processing of speech and speaker information (Chandrasekaran et al., 2011; Kaganovich et al., 2006; Schweinberger et al., 2011; von Kriegstein et al., 2010). Further support for integrated processing comes from psychophysiological research showing that listeners retain information about a speaker's voice in long-term memory (e.g., Palmeri et al., 1993) and use their knowledge about the

vocal characteristics of the speaker to enhance linguistic processing (e.g., Nygaard and Pisoni, 1998).

In this thesis, I provide further empirical evidence that processes involved in the analysis of speech and speaker information interact on the neural and behavioral level. In Study 1, I present data from an experiment which used functional magnetic resonance imaging (fMRI) to show that brain regions sensitive to acoustic speaker parameters are involved in the linguistic processing of speech, and that these regions are functionally connected with speech-sensitive regions when dealing with acoustic differences across speakers during speech recognition. Based on these and previous findings (von Kriegstein et al., 2010), I propose a neural mechanism that exploits functional interactions between speech- and speaker-sensitive areas in left and right hemispheres to allow for robust speech recognition in the context of speaker variations. This mechanism assumes that speech recognition, including recognition of linguistic prosody, predominantly involves areas in the left hemisphere. In Study 2, I present two fMRI experiments that investigated the hemispheric lateralization of linguistic prosody recognition in comparison to recognition of the speech message and speaker identity, respectively. Although linguistic prosody information is often assumed to be predominantly processed in the right hemisphere (e.g., Friederici and Alter, 2004), the results showed a clear left-lateralization when recognition of linguistic prosody was contrasted against speaker recognition. Study 3 used a different approach to show interactions in the processing of speech and speaker information on the behavioral level. Here, I investigated the conditions under which listeners benefit from prior exposure to a speaker's voice in speech recognition. The results suggest that listeners implicitly learn acoustic speaker information during a speech task and use such information to improve comprehension of speech in noise.

In the following chapter, I summarize previous findings on the neural processing of acoustic speaker information. In Chapter 3, I describe the acoustic parameters used in speech and speaker recognition In Chapter 4, I review evidence for different theoretical approaches to speech recognition in the context of speaker variations. Summaries of the empirical studies with a particular focus on their integration with and contribution to existing literature are provided in Chapter 5. Studies 1 and 2 have been published in *NeuroImage*. Links to the published manuscripts are provided in Chapters 6 and 7. The complete manuscript of Study 3 is appended as an individual chapter.

2 Neural Processing of Acoustic Speaker Information

While previous research has studied the neural basis of speech perception in great detail, relatively little research has been performed on the neural processes involved in the perception of acoustic speaker information (reviewed by Belin et al., 2004). In contrast to speech-specific areas which have been predominantly found in the left hemisphere (e.g., Leff et al., 2008; McGettigan and Scott, 2012; Scott, 2005; Vigneau et al., 2006), early clinical studies suggest that lesions to the right hemisphere are associated with impairments in recognizing a person by voice (e.g., Assal et al., 1976; 1981; Van Lancker and Canter, 1982; Van Lancker et al., 1989). The phenomenon of impaired voice recognition in the presence of preserved face recognition has been called 'phonagnosia', by analogy to 'prosopagnosia', which describes the reverse pattern of impairment (Van Lancker and Canter, 1982). More recently, a first case of congenital phonagnosia has been reported (Garrido et al., 2009): patient KH was specifically impaired on her ability to recognize famous persons by voice and to learn previously unknown voices while performing normal on a set of control tasks and showing no sign of structural brain abnormalities which leaves open to future research of how her deficits in acoustic speaker processing relate to neural dysfunction.

In addition to clinical studies, functional neuroimaging has been used to localize brain regions involved in the processing of acoustic speaker information. For example, Belin and colleagues (2000) presented listeners with human vocal and non-vocal environmental sounds while MRI scans were performed. The results showed that regions along superior temporal sulcus (STS) responded more strongly to vocal than non-vocal sounds. Selective responses to vocal sounds were found in both hemispheres, but consistent with lesion studies, they appeared to be stronger in the right hemisphere. Other fMRI studies using similar contrasts confirmed the finding of 'voice-selective' areas in superior temporal regions (e.g., Belin et al., 2002; Fecteau et al., 2004; Gervais et al., 2004). Von Kriegstein and colleagues (2003) used a different approach to localize brain regions involved in the processing of acoustic speaker information. Instead of contrasting conditions that differed in stimulus features, they instructed their participants to perform two different tasks whilst listening to the same stimulus material: a speaker task that required analysis of the speaker's voice; and a speech task that required analysis of the speech message. The results showed stronger involvement

of right anterior STS in the speaker task than the speech task, suggesting that this region is involved in the active analysis of voice information. Although the processing of voice information has been shown to involve various cortical and subcortical areas (e.g., Latinus et al., 2011; Nakamura et al., 2001; von Kriegstein et al., 2006), the majority of previous research suggests a crucial role of temporal areas. In particular, three different areas distributed along temporal cortex have been found to support distinct functions in voice processing: anterior temporal areas have been suggested to process voice identity information, whilst middle and posterior areas have been found to be involved in the acoustic analysis of different speaker parameters (see next chapter) (Andics et al., 2010; Belin and Zatorre, 2003; Formisano et al., 2008; von Kriegstein and Giraud, 2004; von Kriegstein et al., 2007; 2010).

3 Acoustic Parameters in Speech and Speaker Recognition

3.1 Acoustic Parameters in Speaker Recognition

When recognizing a familiar person by voice, we can use a multitude of different cues that are available in the acoustic signal, and each speaker seems to be characterized by a unique set of voice features (Lavner et al., 2000). Listeners can, for instance, describe a voice as breathy, nasal, high, coarse, and thin, to name but a few common attributes (for review, see Kreiman et al., 2005). However, two acoustic parameters in particular have been shown to be of major importance for speaker recognition. The first is the average oscillation rate of the glottal folds or glottal pulse rate (GPR), which determines the fundamental frequency (*f*0) of a speech sound and is perceived as voice pitch. The second is vocal tract length (VTL), which influences the position of formants in the spectral envelope and is perceived as voice timbre.

Evidence for the contribution of GPR to speaker recognition comes from the observation that listeners are able to identify familiar speakers in the complete absence of vocal tract information; that is, when speaker recognition solely relies on glottal fold information (Abberton and Fourcin, 1978). Moreover, previous studies have investigated the perceived similarity between speech samples produced by different speakers (e.g., Baumann and Belin, 2010; Matsumoto et al., 1977; Walden et al., 1978). Using multidimensional scaling analysis, these studies suggest that average GPR is a salient cue in speaker recognition as listeners' similarity ratings strongly relied on similarity across speakers in this parameter.

A different approach to the study of acoustic parameters in speaker recognition is to use sophisticated vocoder software, such as STRAIGHT (Kawahara and Irino, 2004), to manipulate voice features experimentally. Gaudrain and colleagues (2009), for instance, instructed listeners to perform similarity judgments on pairs of speech samples that were identical in voice features or differed by manipulation of either average GPR or VTL. They found that smaller differences in VTL than GPR led to the perception of different speakers indicating that VTL is the more reliable cue in speaker recognition. The authors explain this finding by the fact that VTL is effectively fixed for a given speaker, whereas GPR varies during natural conversation to convey prosody information (see next section).

Of course, we might rely on many more acoustic cues than just average GPR and VTL when recognizing a person by voice. Remez and colleagues (1997) found that listeners are able to recognize speakers from sine-wave replicas of natural speech that preserve no glottal fold and only very little vocal tract information. However, the study of other acoustic speaker parameters has been proven difficult since – in contrast to average GPR and VTL which are closely related to the percept of voice pitch and voice timbre, respectively – most of them lack a clear perceptual counterpart (Maryn et al., 2009).

3.2 Overlap in Acoustic Parameters Relevant to Speech and Speaker Recognition

Glottal fold and vocal tract parameters do not only play an important role in speaker recognition but they are also crucial in determining linguistic aspects of the speech signal. According to the source-filter theory of speech production (e.g., Fant, 1960; Stevens and House, 1961), speech sounds are the product of a two-stage process in which a sound source is generated by the glottal folds and then filtered by the resonant properties of the vocal tract. While average GPR can be used as a cue to voice pitch in speaker recognition, dynamic variations of GPR over multiple speech segments carry prosody information which is not only important for emotional (emotional prosody) but also for linguistic aspects of the signal (linguistic prosody). In tone languages such as Mandarin, linguistic prosody conveys word meaning (e.g., Howie, 1976). In intonational languages such as English and German, linguistic prosody signals, amongst other things, whether a sentence is a question or a statement by rising or falling pitch contours (e.g., Couper-Kuhlen, 1986). Similarly, vocal tract parameters serve as a cue to speaker as well as speech recognition. Vocal tract dynamics are crucial to the production of phonemes which is accomplished by alteration of the filtering properties of the vocal tract which, in turn, is a result of movements of the articulators (e.g., tongue, lips, and jaw).

A dramatic example of how speech- and speaker-specific vocal tract parameters interact in natural speech is given by the classical study of Peterson and Barney (1952). In this study, a total of 76 American English speakers were recorded while producing monosyllabic words with different vowels of the form /h-vowel-d/. Using sound spectrography, the frequencies of the first two vowel formants were measured. The results showed a considerable overlap in the

formant space covered by the different vowels indicating that a particular vowel spoken by one speaker could be acoustically similar to a different vowel spoken by another speaker (e.g., one speaker's *hid* could be more similar to another speaker's *head*). The finding of speaker-related variations in vowel formant space has been since then repeatedly replicated and it has been shown that acoustic differences across speakers are even larger when speakers with different regional dialects are included (Clopper et al., 2005; Hagiwara, 1995; Hillenbrand et al., 1994).

In order to correctly identify vowels under such variable conditions, listeners have to map many different acoustic patterns to the same phonemic category. Nevertheless, listeners have little difficulty with vowel identification in the context of acoustic speaker variations (Abramson and Cooper, 1959; Peterson and Barney, 1952). Theoretical approaches, however, have struggled to produce a widely accepted explanation for this remarkable robustness to speaker variations (cf. Nusbaum and Magnuson, 1997; Pisoni, 1997; reviewed in the next chapter), and automatic speech recognition devices perform well below normal hearing listeners when dealing with acoustic differences across speakers (O'Shaughnessy, 2008).

4 Speech Recognition in the Context of Acoustic Speaker Variations

4.1 Theoretical Approaches

The vast majority of research on speech recognition has been performed in isolation from research on speaker recognition, and speaker-related variations have been long regarded as a source of noise in the speech signal (reviewed by e.g., Nygaard, 2005). Under the abstractionist view, for instance, robust speech recognition relies on a process of perceptual normalization in which speaker-related variations are filtered from the signal to arrive at abstract canonical linguistic units (reviewed by Pisoni, 1997).

It should be noted that not all normalization theories assume that speaker information needs to be filtered out or discarded to allow for robust speech recognition. An important distinction is as to whether a process of intrinsic or extrinsic normalization is assumed (Ainsworth, 1975; Magnuson and Nusbaum, 2007; Nearey, 1989): while intrinsic normalization implies that any given speech sample contains sufficient information to normalize for speaker variations (e.g., Shankweiler et al., 1977; Syrdal and Gopal, 1986), extrinsic normalization theories assume that listeners also use prior information about the speaker to account for speaker variations (e.g., Joos, 1948). The latter is supported by the finding that prior context information can affect the perception of linguistic categories. In their seminal study, Ladefoged and Broadbent (1957) presented listeners with monosyllabic test words which followed a standard sentence (*Please say what word this is*.). Importantly, different versions of this sentence were synthesized by shifting formant frequencies, thereby simulating speakers with different vocal tract characteristics. Listeners' perception of physically identical test words was strongly influenced by these formant shifts (e.g., the same test word was perceived as *bit* following one version of the standard sentence and as *bet* following another version in which formants were shifted towards lower frequencies). This suggests that speaker information – rather than simply being noise that needs to be filtered from the signal – is actively exploited during linguistic analysis to shape the interpretation of subsequent speech samples. This is also consistent with the idea of active control mechanisms that allow for adaptation to speaker-related variations (Nusbaum and Magnuson, 1997) and with the finding that understanding

speech from multiple as opposed to single speakers is associated with increased processing demands (e.g., Magnuson and Nusbaum, 2007).

A radically different approach to the problem of speaker-related variations in speech recognition has been proposed by exemplar-based models (e.g., Goldinger, 1998). Based on findings that suggest common memory traces for speech and speaker information (e.g., Goldinger et al., 1991; Palmeri et al., 1993; Pisoni, 1993), these models assume that information about the vocal characteristics of a speaker is represented in long-term memory together with linguistic information by instance-specific exemplars. Since detailed information about the entire speech event, including speaker information, is preserved, exemplar-based models can explain robust speech recognition without the need for speaker normalization.

4.2 Segregated vs. Integrated Processing of Speech and Speaker Information

Despite the variety of theoretical approaches to speaker-related variations in speech recognition, one can generally categorize theories with regard to whether they assume a segregated or integrated processing of speech and speaker information. This reduces the full range of theoretical considerations to a simple dichotomy which might prove useful for deriving testable hypotheses about the neural basis of speech recognition in the context of speaker variations.

Figure 4-1A shows an example of segregation in the processing of speech and speaker information that is largely compatible with the abstractionist view. This model assumes that, after early structural analysis of the auditory input, information relevant to speech and speaker recognition is processed in segregated, non-interacting pathways with a module for vocal speech analysis in the left hemisphere and a module for voice identity analysis in the right hemisphere. As a consequence, acoustic features specific to a speaker's voice are discarded at early processing stages and not used during a stage of vocal speech analysis – consistent with a process of speaker normalization as suggested by abstractionist models (reviewed by Pisoni, 1997).

Findings from a recent fMRI study have been interpreted as support for an early process of speaker normalization (Salvata et al., 2012). Using an fMRI adaptation paradigm, this study found greater sensitivity to phoneme changes (when spoken by the same and different speakers) than to constant phonemic input (when spoken by the same and different speakers) in bilateral anterior portions of middle superior temporal gyrus (STG). However, the claim that these regions are involved in an early process of speaker normalization is questionable because they appear to be in close vicinity to anterior temporal regions that have been found to show sensitivity to voice identity information (e.g., Andics et al., 2010; see Chapter 2) and therefore represent good candidate areas for voice identity analysis. Furthermore, superior temporal areas have been found to be sensitive to vocal stimuli (e.g., Belin et al., 2000) and involved in the active processing of voice information (e.g., von Kriegstein et al., 2003). The proposed model of segregated speech and speaker processing, by contrast, implicitly assumes that speaker normalization takes place earlier in the processing hierarchy; that is, before speaker information is actively analyzed. Another fMRI study on the neural basis of speaker normalization found that a network of temporal and parietal areas is involved in speech recognition when dealing with speaker variations as compared to speech recognition under constant speaker conditions (Wong et al., 2004). This has been, however, taken as evidence for the existence of active control mechanisms (Magnuson and Nusbaum, 2007) which are not compatible with a model of segregated processing of speech and speaker information and processes of speaker normalization as suggested by abstractionist models.

An alternative view is that speech and speaker information is processed in an integrated fashion (Fig. 4-1B). Similar to the model of segregated processing, this view assumes an initial structural analysis of the auditory input and separate modules for vocal speech and voice identity analyses in left and right hemispheres, respectively. The critical difference is, however, that processes involved in the analyses of speech and speaker information interact via forward, backward, and lateral connections. This implies that, during linguistic processing, information about the speaker is still available and can be actively used to improve speech recognition in the context of speaker variations.

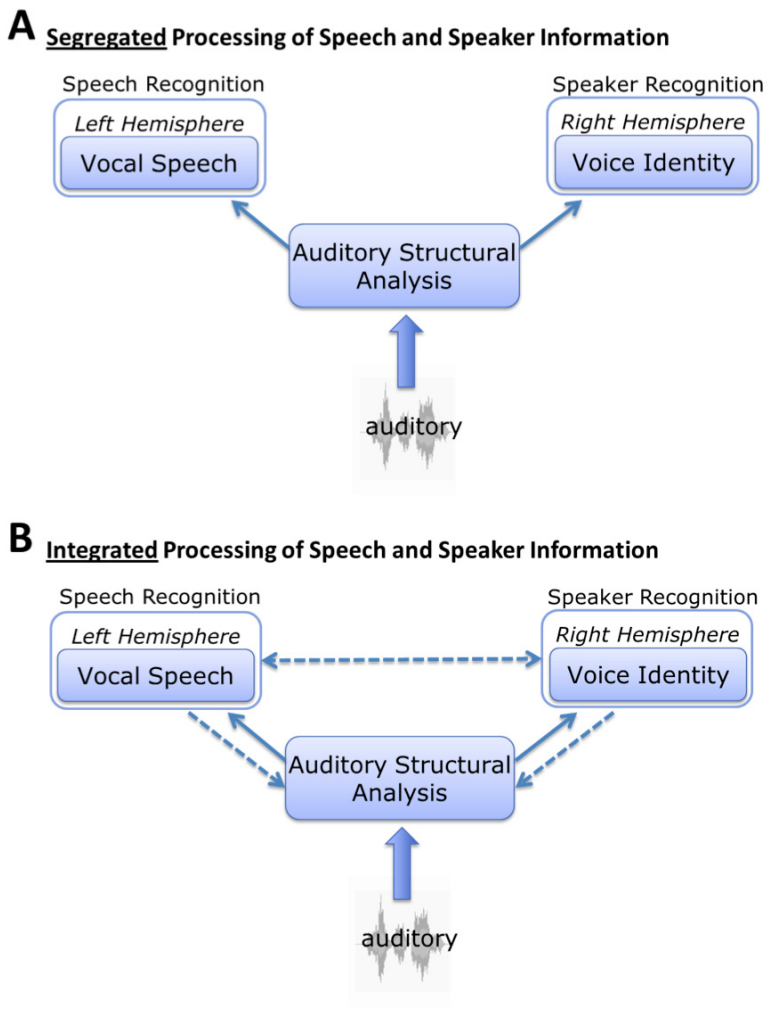


Figure 4-1. Models of segregated (A) and integrated (B) processing of speech and speaker information. **A.** This model assumes that speech and speaker information is processed in a *segregated* fashion. After early structural analysis, information relevant to speech and speaker recognition is processed separately with a module for vocal speech analysis in the left hemisphere and a module for voice identity analysis in the right hemisphere. **B.** An alternative view assumes that speech and speaker information is processed in an *integrated* fashion. Speech- and speaker-related processes interact via forward (solid arrows), backward and lateral connections (dashed arrows).

Such a model has been previously described as a hybrid between abstractionist and exemplar-based models because it exploits both the extraction of abstract linguistic features and the representation of speaker information during speech recognition (von Kriegstein et al., 2010). The same study also provides evidence that the brain processes speech and speaker information in an integrated rather than segregated fashion when both types of information are determined by vocal tract parameters. Specifically, it was found that areas in right posterior STG/STS, which are sensitive to speaker-specific vocal tract parameters (i.e., VTL), are involved in speech recognition when speakers concomitantly changed in VTL. Furthermore, these VTL-sensitive areas were functionally connected to homologous areas in the left hemisphere, which are involved in speech recognition, when dealing with VTL-induced speaker changes during speech recognition. Further support for an integrated processing of speech and speaker information comes from another fMRI study showing that left posterior middle temporal gyrus responds to changes in both speech message and speaker identity (Chandrasekaran et al., 2011). In addition, electroencephalography (EEG) studies

found interactions in the cortical processing of speech and speaker information (Kaganovich et al., 2006; Schweinberger et al., 2011). At the behavioral level, it has been shown that unimpaired but not dyslexic listeners are better able to recognize a speaker when presented in their native as opposed to a foreign language suggesting that speaker recognition depends on the ability to process the input linguistically (Perrachione et al., 2011). That interactions in the processing of speech and speaker information also occur in the other direction has been demonstrated by the finding that familiarity with a speaker's voice can enhance speech recognition (Levi et al., 2011; Magnuson et al., 1995; Nygaard et al., 1994; Nygaard and Pisoni, 1998; Yonan and Sommers, 2000). Taken together, there is increasing evidence from behavioral and neuroimaging studies for an integrated processing of speech and speaker information which challenges the biological plausibility of intrinsic speaker normalization and abstractionist models.

5 Summary of Empirical Studies

5.1 Study 1

5.1.1 Research Aim

Although the problem of speaker-related variations in speech recognition has been approached from various theoretical perspectives, an integrated processing of speech and speaker information might provide the neural basis for robust speech recognition when dealing with acoustic differences across speakers (see previous chapter). In particular, a recent fMRI study suggests that the ability to understand speech from different speakers relies on functional interactions between areas in the left and right hemispheres that are sensitive to specific acoustic features of speech and speaker (von Kriegstein et al., 2010). Such interactions were found for vocal tract parameters, which determine parts of the speech message and the timbre of a speaker's voice. It remained unclear whether interactions between specific areas in the two hemispheres are restricted to speech- and speaker-specific vocal tract parameters or whether they represent a general feature of how the brain accomplishes speech recognition when dealing with speakers variations. Study 1 exploited the property of glottal fold parameters to determine both speech and speaker information (i.e., linguistic prosody for speech and voice pitch for speaker; see Chapter 3) to investigate whether a similar interaction exists for speech- and speaker-specific GPR.

5.1.2 Study Summary

Participants of Study 1 were presented with sequences of syllables and were instructed to recognize linguistic prosody from speakers who differed in average GPR. Blood-oxygen-level dependent (BOLD) responses were measured in silence periods between two syllable sequences using fMRI. Linguistic prosody was defined by dynamic variations in GPR (i.e., rising and falling pitch contours typical for the intonation of questions and statements, respectively). Importantly, modern vocoder software (Kawahara et al., 2008) was used to ensure that information relevant to linguistic prosody as well as speaker identity was solely based on GPR variations. This means that linguistic prosody and speaker information was fundamentally intermingled in the same acoustic parameter (i.e., GPR) and that participants had to disentangle speech- from speaker-specific GPR when recognizing linguistic prosody from speakers who differed in GPR. The experiment also included several control conditions,

namely syllable sequences in which speakers varied in VTL rather than GPR, and a control task that involved speaker recognition. The experiment therefore had a 2×2 factorial design with the factors task (prosody vs. speaker task) and speaker change (GPR change vs. VTL change).

The first hypothesis was that regions in right Heschl's gyrus, which are sensitive to speaker-specific GPR, are involved in recognizing linguistic prosody from speakers who differed in GPR. The second hypothesis was that right Heschl's gyrus is functionally connected to its homologous area in the left hemisphere during this process. These hypotheses mirrored previous findings on the processing of speech- and speaker-specific vocal tract parameters (von Kriegstein et al., 2010). However, for GPR processing, an involvement of Heschl's gyrus rather than posterior STG/STS was expected (e.g., Griffiths et al., 2010; Kumar et al., 2011; Wong et al., 2008).

To test the first hypothesis, BOLD responses elicited by the 'task × speaker change interaction' ([prosody task/GPR change > speaker task/GPR change] > [prosody task/VTL change > speaker task/VTL change]) were analyzed. This interaction analysis ensured that the observed BOLD response was specific to the recognition of linguistic prosody from speakers who differed in GPR. The 'VTL change' condition and the speaker task were employed to control for the possibility that BOLD responses reflect a general increase in activity only due to GPR-induced speaker changes or only due to the prosody task. The results supported the first hypothesis by showing that posteromedial and central portions of right Heschl's gyrus (i.e., TE1.1 and TE1.0; Morosan et al., 2001) are involved in recognizing linguistic prosody from speakers who differed in GPR.

A psychophysiological interaction (PPI) analysis (Friston et al., 1997) was used to address the second hypothesis. Here, the seed region was set to right Heschl's gyrus (based on the results of the interaction analysis); the target region was the homologous area in the left hemisphere. To ensure that a functional connection between left and right Heschl's gyri was specific to the recognition of linguistic prosody in the context of GPR-induced speaker changes, the psychological variable was defined by the 'task × speaker change interaction' ([prosody task/GPR change > speaker task/GPR change] > [prosody task/VTL change > speaker task/VTL change]). The results supported the second hypothesis by showing an increased functional connection between right Heschl's gyrus and the posteromedial part of

left Heschl's gyrus (TE1.1) during the recognition of linguistic prosody when speakers differed in GPR.

5.1.3 Conclusions

The results of Study 1 were difficult to reconcile with the view of a segregated processing of speech and speaker information, but they were in full agreement with an integrated-processing view. In combination with previous findings on vocal tract parameters (von Kriegstein et al., 2010), the results suggest that robust speech recognition in the context of acoustic speaker variations is based on a neural mechanism that exploits functional interactions between areas sensitive to specific acoustic features of speech and speaker. Acoustic features specific to a speaker's voice might be extracted in specialized areas in the right hemisphere and communicated to areas involved in the analysis of speech-specific acoustic features in the left hemisphere to improve speech recognition when dealing with speaker variations. Such a mechanism is consistent with active processing of speaker information during speech recognition (e.g., Magnson and Nusbaum, 2007) and capable of explaining behavioral findings showing that knowledge of the vocal characteristics of a speaker enhances linguistic processing (e.g., Nygaard et al., 1994; Nygaard and Pisoni, 1998).

5.2 Study 2

5.2.1 Research Aim

Implicit in the neural mechanism proposed in Study 1 is the assumption that speech recognition, including recognition of the speech message as well as linguistic prosody, is supported by functions of the left hemisphere. While most speech-related processes, such as lexical and syntactic processes, have been found to predominantly involve left-hemispheric brain regions, linguistic prosody information has been assumed to be predominantly processed in the right hemisphere (for review, see e.g., Friederici, 2011). This has been explained by a right-hemispheric superiority in processing suprasegmental input (Friederici and Alter, 2004); that is, information relevant to linguistic prosody usually varies over multiple speech segments, whereas phoneme recognition, for example, is based on segmental analysis of the input. However, studies investigating the neural processing of linguistic as compared to emotional prosody information challenge the assumption of right-hemispheric

lateralization in linguistic prosody recognition (Wildgruber et al., 2004; Witteman et al., 2011). Furthermore, a left-hemispheric involvement has been found when linguistic prosody signals word meaning in tone languages (Gandour et al., 2000). The aim of Study 2 was therefore to clarify the roles of left and right hemispheres in linguistic prosody recognition and to investigate the neural substrates of linguistic prosody in comparison to speaker recognition, which are currently unknown.

5.2.2 Study Summary

Study 2 comprised two fMRI experiments. Participants of the first experiment were presented with sequences of short German sentences that differed in a single phoneme (e.g., *Er schreibt*. vs. *Er schreit.*) and in linguistic prosody (i.e., sentences were intonated as questions and statements). On these sequences, participants performed either a prosody task that involved recognition of linguistic prosody or a speech task that involved recognition of the speech message. BOLD responses were measured in silence periods between two sentence sequences using fMRI. Similar to the experiment of Study 1, this experiment additionally comprised two types of speaker changes (GPR change vs. VTL change). This was done to increase similarity to the second experiment of this study. In summary, the first experiment had a 2×2 factorial design with the factors task (prosody vs. speech task) and speaker change (GPR change vs. VTL change). The second experiment was the same as the fMRI experiment used in Study 1; that is, participants were presented with sequences of syllables that differed in linguistic prosody as well as in acoustic speaker parameters (GPR change vs. VTL change) and were asked to recognize either linguistic prosody (prosody task) or speaker identity (speaker task). However, instead of investigating the interaction between task and speaker change, the analyses of Study 2 focused on BOLD responses elicited by the main effects of prosody task (i.e., [prosody task > speech task] in Experiment 1, and [prosody task > speaker task] in Experiment 2).

Based on previous literature, it was expected that recognition of linguistic prosody in both experiments would involve language-relevant areas in temporal and frontal lobes (Friederici, 2002, 2011; Hickok and Poeppel, 2007; Vigneau et al., 2006). It was further expected that BOLD responses elicited by the main effect of prosody task in Experiment 1 would predominantly involve right-hemispheric areas since the contrast of prosody vs. speech task particularly focused on the analysis of suprasegmental information (Friederici and Alter, 2004). As the neural substrates of prosody in comparison to speaker recognition are currently

unknown, there were two opposing hypotheses concerning lateralization of linguistic prosody recognition in Experiment 2: either prosody and speaker recognition are supported by different areas in the right hemisphere in which case BOLD responses in the contrast of prosody vs. speaker task might be right-lateralized or the neural substrates of prosody and speaker recognition overlap in the right hemisphere which could result in a predominant left-lateralization of prosody as compared to speaker recognition.

The results showed that recognition of linguistic prosody in both experiments involved a large network of brain regions, including language-relevant areas in temporal and frontal lobes. As expected, prosody as compared to speech recognition (i.e., main effect of prosody task in Experiment 1) predominantly involved temporo-frontal areas in the right hemisphere. When prosody was contrasted against speaker recognition (i.e., main effect of prosody task in Experiment 2), however, activation in temporo-frontal areas was predominantly lateralized to the left hemisphere. A conjunction analysis ([prosody task > speech task] ∩ [prosody task > speaker task)]) revealed that left supramarginal gyrus as well as bilateral inferior frontal gyri were commonly activated by the prosody tasks of both experiments (i.e., irrespective of whether the prosody task was contrasted against speech or speaker task). The finding that the neural substrates of linguistic prosody recognition were differentially lateralized in the two experiments was confirmed by additional laterality analyses. A series of control analyses showed that this differential lateralization cannot be explained by differences in stimuli or prosody tasks between the experiments.

5.2.3 Conclusions

The results of Study 2 showed that linguistic prosody information is processed in both hemispheres with hemispheric lateralization depending on whether linguistic prosody was contrasted against speech or speaker recognition. A right-hemispheric lateralization for prosody as compared to speech recognition is consistent with findings from previous studies that manipulated the degree to which segmental phonemic and suprasegmental prosodic information was available in the stimuli (e.g., Meyer et al., 2002; Plante et al., 2002). The present findings additionally showed that right-hemispheric areas do not only respond to suprasegmental information but are also involved in its active analysis during prosody recognition. A stronger involvement of left-hemispheric areas in linguistic prosody as compared to speaker recognition is consistent with the functional lateralization hypothesis which suggests a left-hemispheric involvement in the analysis of linguistically relevant pitch

information (Van Lancker, 1980). Furthermore, this finding provides support for a neural mechanism (proposed in Study 1) in which the processing of speech and speaker information relies on functions of left and right hemispheres, respectively. Taken together, the results of Study 2 suggest that recognition of linguistic prosody is based on a dynamic hemispheric interplay that captures both the right-hemispheric advantage in processing suprasegmental pitch information and the left-hemispheric dominance in speech processing.

5.3 Study 3

5.3.1 Research Aim

While the first two studies included experimental manipulations that involved rapid changes in acoustic speaker parameters, Study 3 used a different approach to gain insight into the processing of speech and speaker information: instead of probing mechanisms underlying robust speech recognition in the context of speaker variations, Study 3 investigated under which conditions listeners benefit from familiarity with a speaker's voice when understanding speech in noise. Previous research showed that listeners are better able to understand speech when it is spoken by a familiar as opposed to an unfamiliar speaker (Levi et al.,, 2011; Magnuson et al., 1995; Nygaard et al., 1994; Nygaard and Pisoni, 1998; Yonan and Sommers, 2000). Typically, these studies employed a paradigm in which listeners were trained to identify a set of speakers by voice-name associations (e.g., Burk et al., 2006, Nygaard et al., 1994; Nygaard and Pisoni, 1998; but see Yonan and Sommers, 2000). Effects of speaker familiarity were then assessed by testing listeners' ability to understand speech from those same speakers (i.e., familiar speakers) and a set of novel speakers (i.e., unfamiliar speakers). Study 3 was designed to investigate whether listeners also benefit from prior exposure to a speaker's voice without explicit speaker training; that is, when the training requires listeners to focus on the speech message rather than speaker identity.

5.3.2 Study Summary

Participants of Study 3 performed an experiment that comprised five sessions conducted on five consecutive days. Over the course of the entire experiment, participants heard sentences similar to those used in the first experiment of Study 2. Sentences were overlaid with speech-shaped noise and participants performed a speech recognition task while speech-reception

thresholds (SRTs) were measured using an adaptive tracking procedure (Kaernbach, 1991). During the training phase which was conducted on the first four sessions of the experiment, participants were presented with speech from one of four male speakers (i.e., familiar speaker)[1]. The choice of familiar speaker was counterbalanced across participants. The test session was conducted on the fifth day of the experiment. At test, participants heard novel sentences from the familiar speaker and three novel speakers (i.e., unfamiliar speakers). Half of the participants performed a test session in which the speaker was blocked over a series of sentences (blocked paradigm). The other half performed a test session in which speakers randomly changed across sentences (interleaved paradigm). In summary, the experiment had a 2×2×4 factorial design with the within-subject factor familiarity (familiar vs. unfamiliar speaker), and the between-subject factors paradigm (blocked vs. interleaved) and familiar speaker (i.e., the speaker a given participant heard during training; speaker 1–4).

Importantly, both training and test sessions required speech recognition; that is, at no time of the experiment, participants were asked to attend to speaker identity. If participants benefited from such *implicit* speaker training, they should be better able to understand speech in noise from familiar than unfamiliar speakers at test. To test this, a three-way mixed-design ANOVA with all design factors was performed on SRTs in the test session. The results showed a significant main effect of familiarity indicating that participants recognized speech from familiar speakers at lower SRTs than speech from unfamiliar speakers ($p = 0.035$). This means that participants were better able to understand speech in noise when it was spoken by the same speaker they heard during training than when spoken by novel speakers. The results also showed a significant interaction between familiarity and familiar speaker ($p < 0.001$) indicating that effects of speaker familiarity depended on which specific speaker the participants heard during training. Post-hoc tests revealed that benefits from speaker familiarity were largest for those speakers for which lowest SRTs during the training session were found. This suggests that benefits from speaker familiarity, at least, partly depended on the intelligibility of the speaker participants heard during training. Additional analyses showed that the differences in speaker intelligibility cannot be explained by differences in energetic masking of noise across speakers or speaker-specific $f0$ range, which has been suggested to be associated with speaker intelligibility (Bradlow et al., 1996).

[1]Note that in the manuscript of Study 3 (Chapter 8) the term *speaker* was replaced by *talker* due to common nomenclature in the literature (e.g., Magnuson et al., 1995; Nygaard and Pisoni, 1998; Levi et al., 2011).

5.3.3 Conclusions

The results of Study 3 provide further support for an integrated processing of speech and speaker information. They showed that familiarity with a speaker's voice enhances comprehension of speech in noise. In contrast to previous studies (Levi et al., 2011; Magnuson et al., 1995; Nygaard et al., 1994; Nygaard and Pisoni, 1998), familiarity was not accomplished by explicit speaker training but induced implicitly during speech recognition. That participants benefited from such incidental speaker exposure suggests that listeners implicitly learn acoustic speaker information during linguistic analysis of the speech signal and exploit this information to improve speech recognition. It should be, however, noted that several limitations to the effect of speaker familiarity on speech recognition have been described in previous literature: Magnuson and colleagues (1995) found that listeners only benefited from personally familiar speakers but not when listeners were familiarized with previously unknown speakers during the experiment. Nygaard and Pisoni (1998) showed that only listeners who successfully learned to identify speakers benefited from speaker familiarity. Furthermore, effects of speaker familiarity have been found to depend on the context in which speakers are learned (Levi et al., 2011; Nygaard and Pisoni, 1998). The results of Study 3 additionally suggest that effects of speaker familiarity depend on speaker intelligibility.

6 Published Manuscript of Study 1

A Neural Mechanism for Recognizing Speech Spoken by Different Speakers

Jens Kreitewolf[1], Etienne Gaudrain[2,3] and Katharina von Kriegstein[1,4]

[1]Max Planck Institute for Human Cognitive and Brain Sciences, Max Planck Research Group Neural Mechanisms of Human Communication, D-04103 Leipzig, Germany

[2]University of Groningen, University Medical Center Groningen, Department of Otorhinolaryngology/Head and Neck Surgery, 9700 RB Groningen, Netherlands

[3]University of Groningen, Graduate School of Medical Sciences, Research School of Behavioural and Cognitive Neurosciences, 9713 GZ Groningen, Netherlands

[4]Humboldt University of Berlin, Psychology Department, D-12489 Berlin, Germany

Published in:

Kreitewolf, J., Gaudrain, E., von Kriegstein, K., 2014. A neural mechanism for recognizing speech spoken by different speakers. NeuroImage, 91, 375-385.

doi:10.1016/j.neuroimage.2014.01.005

7 Published Manuscript of Study 2

Hemispheric Lateralization of Linguistic Prosody Recognition in Comparison to Speech and Speaker Recognition

---

Jens Kreitewolf[1], Angela D. Friederici[2] and Katharina von Kriegstein[1,3]

[1]Max Planck Institute for Human Cognitive and Brain Sciences, Max Planck Research Group Neural Mechanisms of Human Communication, D-04103 Leipzig, Germany

[2]Max Planck Institute for Human Cognitive and Brain Sciences, Department of Neuropsychology, D-04103 Leipzig, Germany

[3]Humboldt University of Berlin, Psychology Department, D-12489 Berlin, Germany

Published in:

Kreitewolf, J., Friederici, A.D., von Kriegstein, K., 2014. Hemispheric lateralization of linguistic prosody recognition in comparison to speech and speaker recognition. NeuroImage, 102, 332-344.

doi:10.1016/j.neuroimage.2014.07.038

8 Manuscript of Study 3

Effects of Talker Familiarity on the Comprehension of Auditory Speech in Noise

Jens Kreitewolf[1], Samuel R. Mathias[2] and Katharina von Kriegstein[1,3]

[1]Max Planck Institute for Human Cognitive and Brain Sciences, Max Planck Research Group Neural Mechanisms of Human Communication, D-04103 Leipzig, Germany

[2]Boston University, Center for Computational Neuroscience and Neural Technology, Boston, MA 02215, USA

[3]Humboldt University of Berlin, Psychology Department, D-12489 Berlin, Germany

Abstract

The human auditory system shows a remarkable robustness to acoustic differences across talkers that introduce large amounts of variability into the speech signal. Previous work has shown that prior knowledge about the vocal characteristics of a talker helps to resolve some of this variability and that familiarity with a talker's voice enhances linguistic processing. Typically, talker familiarity was induced by explicit voice training in previous studies. Here, we tested whether listeners also benefit from implicit voice training by using a training paradigm in which listeners were familiarized with a talker's voice incidentally while recognizing speech in noise. During four training sessions, listeners heard short sentences spoken by one talker. In the final test session, listeners heard novel sentences spoken by the same talker who was presented during training (familiar talker) and three novel talkers (unfamiliar talkers). For half of the listeners, the talker was blocked over a series of sentences (blocked paradigm). For the other half, talkers randomly changed across sentences (interleaved paradigm). In both training and test sessions, listeners were asked to identify the verb of the currently presented sentence while speech-reception thresholds (SRTs) were measured using an adaptive tracking procedure. The results showed that the listeners were better able to understand speech in noise when it was produced by the familiar talker than the unfamiliar talkers. However, this 'familiarity benefit' depended on which specific talker the listeners heard during training. Furthermore, blocked talker presentation at test did not affect the degree to which listeners benefited from talker familiarity. Importantly, our results showed that listeners can benefit from talker familiarity without explicit voice training. This is in accordance with the notion of integrated processing of speech- and talker-specific information and suggests that listeners implicitly learn acoustic talker information during a linguistic task and exploit such information to improve comprehension of speech in noise.

8.1 Introduction

Acoustic speech signals provide the listener with a wealth of information, not only with respect to what is said but also with respect to who is speaking. In speech research, acoustic features specific to a talker's voice are often considered to be a major source of variability (reviewed by Nygaard, 2005) because differences across talkers can result in very large differences in the acoustic waveform, even when the underlying speech message is the same (Peterson and Barney, 1952). One of the biggest challenges faced by the auditory system is to compensate for these acoustic variations across talkers. Previous research suggests that the ability to understand speech in the context of talker variability relies on active processing of talker information during speech recognition (reviewed by Nusbaum and Magnuson, 1997). Furthermore, it has been suggested that much of the variability introduced by acoustic differences across talkers is perceived and memorized along with the speech message (Bradlow et al., 1999; Palmeri et al., 1993; Pisoni, 1993). Variations in talker characteristics, however, clearly have an impact on speech recognition performance. A number of studies showed that recognizing speech from changing talkers is less accurate and can result in longer processing times than recognizing speech under constant talker conditions (e.g., Creelman, 1957; Mullennix et al., 1989; Nusbaum and Morin, 1992).

An important factor in understanding speech from different talkers is whether the talker's voice is familiar to the listener. Prior knowledge about the vocal characteristics of the talker supports adaptation to the talker's voice which, in turn, helps understanding the speech message (Levi et al., 2011; Magnuson et al., 1995; Nygaard et al., 1994; Nygaard and Pisoni, 1998; Yonan and Sommers, 2000). For example, Nygaard and Pisoni (1998) showed that familiarity with a talker's voice improves linguistic processing of speech from that talker. Specifically, they found that a group of listeners which was trained to identify a set of talkers was better able to understand speech from those same talkers than a second group of listeners that did not have prior experience with the talkers. However, the benefit in recognizing speech from the familiar talkers (henceforth referred to as the 'familiarity benefit') depended strongly on the type of stimuli heard during the training and test sessions: learning to identify talkers from sentence-length utterances did not improve the recognition of isolated words spoken by the same talkers. The authors explained this finding by differences in the acoustic talker information contained within sentences and words and suggest that the integration of talker and speech information is stimulus-dependent. Furthermore, it has been shown that the

familiarity benefit also depends on the language context in which talkers are learned (Levi et al., 2011): listeners only benefited from talker familiarity when speech recognition at test was performed in the same language in which talkers have been previously learned. Moreover, these studies demonstrate that only listeners who successfully learned to identify talkers showed improved speech recognition. Another study found that listeners are better able to understand speech from personally familiar talkers than from talkers they were familiarized with during the experiment (Magnuson et al., 1995). Taken together, these findings showed that, under some conditions, listeners can benefit from talker familiarity during speech recognition but that the familiarity benefit depends on the type of acoustic talker information listeners have learned and that mere exposure to a talker's voice does not necessarily result in enhanced speech recognition.

The aim of the present study was to further investigate the conditions under which listeners benefit from talker familiarity during speech recognition. Specifically, we tested whether explicit attention to the talker dimension during training is a necessary prerequisite for the familiarity benefit or whether listeners also benefit from prior exposure to a talker's voice during a speech recognition task. Previous studies on the familiarity benefit induced familiarity using a training paradigm in which listeners' attention was explicitly directed to the talker dimension (e.g., Levi et al., 2011; Magnuson et al., 1995; Nygaard et al., 1994; Nygaard and Pisoni, 1998). Here, we used task demands that required listeners to focus on the speech message rather than talker identity at any time of the experiment. In other words, talker familiarity was induced incidentally during speech recognition without explicit talker learning. This approach might be more similar to talker learning during natural conversation than explicitly directing listeners' attention to acoustic talker features.

The experiment comprised five sessions that were conducted on five consecutive days. On each day, listeners heard short sentences that were overlaid with speech-shaped noise. The task was to listen to each sentence and to select the verb that was present in the sentence from twenty options displayed on a computer screen (Fig. 8-1) while speech-reception thresholds (SRTs) were measured using an adaptive tracking procedure (Kaernbach, 1991). During the first four days of the experiment (training phase), listeners were presented with sentences from a single talker. The fifth day served as test where listeners were presented with novel sentences from the same talker who was presented during training (familiar talker) and three novel talkers (unfamiliar talkers). Note that both training and test phase involved speech recognition and that no explicit voice learning was included. If listeners benefit from training,

they should be better able to understand speech from familiar than from unfamiliar talkers; therefore, participants should attain lower thresholds for familiar than unfamiliar talkers during the final test session.

We also investigated whether a potential familiarity benefit depends on voice continuity. Half of the listeners performed the test session in which talkers were presented in separate blocks of trials (blocked paradigm). For the other half, the test session comprised blocks of trials in which speech from all four talkers (one familiar, three unfamiliar) was randomly intermingled (interleaved paradigm). Listeners in the blocked paradigm can adjust to the talker's voice over a series of utterances, whereas the interleaved paradigm forces listeners to constantly readjust to the current talker. Based on previous findings showing that voice continuity enhances linguistic processing (e.g., Bent and Holt, 2013; Best et al., 2008; Bradlow and Pisoni, 1999; Kitterick et al., 2010), one might expect that listeners would benefit more greatly from talker familiarity in the blocked than in the interleaved condition.

8.2 Methods

8.2.1 Participants

Twenty-four volunteer listeners [16 female; mean age 25.6 years; age range 21-30 years; all right-handed as assessed with the Edinburgh questionnaire (Oldfield, 1971)] participated in the study. All of the listeners were native German speakers. None of them had prior experience with the talkers' voices used in this study. None of them had any history of neurological or psychiatric disorder. All twenty-four participants of the study had hearing levels within normal limits. Hearing levels in both ears were measured for frequencies between 250 and 8000 Hz in octave steps using pure-tone audiometry. Normal hearing was defined as a maximum of 20 dB hearing level (HL) at all frequencies. For one listener, hearing level was 25 dB HL for 2000 Hz, 40 dB HL for 4000 Hz, 50 dB HL for 8000 Hz (all in the left ear), and 30 dB HL for 8000 Hz in the right ear. This listener was excluded from the study prior to the start of the experiment. Written informed consent was collected from all participants according to procedures approved by the Research Ethics Committee of the University of Leipzig. Participants were paid after completing the experiment. This included

an hourly rate for their participation (7 EUR) as well as an additional amount based on their performance on each day of the experiment (2 EUR).

8.2.2 Stimuli

The stimuli used in this study comprised 200 short German sentences, each consisting of one noun and one verb. Each sentence started with *Er* (English: he). The verbs were selected to form pairs of dense lexical neighbors; that is, the members of a given pair differed in a single phoneme only. Half of these pairs differed in consonants (e.g., *Er liebt*. vs. *Er liest*.; English: He loves. vs. He reads.), the other half differed in vowels (e.g., *Er setzt*. vs. *Er sitzt*.; English: He sets. vs. He sits.). Each of four native German talkers (all male; mean age 26.8 years; age range 23-31 years; average *f*0, see Suppl. Tab. 8-1) produced the complete set of sentences. Recordings were made in a sound-attenuating chamber (IAC – I200 series, Winchester, UK) with a resolution of 16 bits and at a sampling rate of 44.1 kHz using a cardioid condenser microphone (RØDE NT55, Silverwater, Australia). After recording, all stimuli were adjusted to the same root mean square (RMS) value using MATLAB (version 7.11, MathWorks, USA). During the experiment, speech stimuli were overlaid with speech-shaped noise which was composed of white noise filtered at the spectrum of the speech sounds. Filtering was accomplished using the fftfilt function as implemented in the MATLAB signal processing toolbox. Speech-shaped noise was generated and mixed with speech stimuli on each trial separately in order to match speech and noise sounds in duration and to present the noise-overlaid speech stimuli at a given signal-to-noise ratio (SNR) (see below). This was done by manipulating the sound level of speech stimuli; the noise sound level was kept constant. Sounds were delivered diotically through headphones (Sennheiser HD580, Wedemark, Germany) using a 16-bit digital-to-analog converter (Creative Sound Blaster Audigy 2 ZS, Jurong East, Singapore) at a sampling rate of 44.1 kHz and a pre-amplifier (Pro-Ject Head Box II, Vienna, Austria).

8.2.3 Procedure

Prior to the experiment, individual hearing levels were measured using a screening audiometer (Micromate 304, Madsen Electronics, Denmark). The experiment included three phases (i.e., *familiarization*, *training*, and *test*) and was conducted on five consecutive days. A graphical summary of the experimental procedure is provided in Figure 8-1.

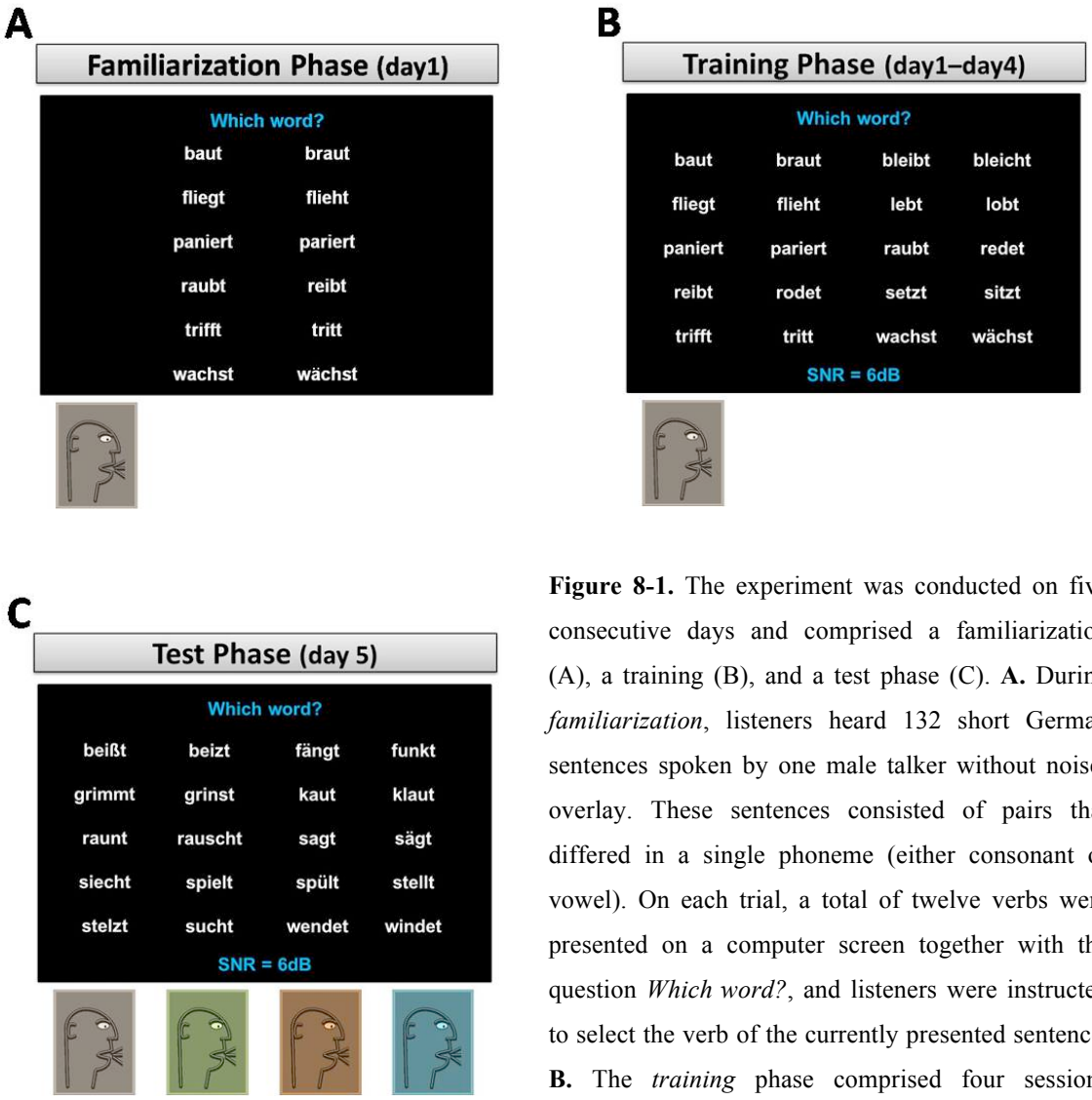


Figure 8-1. The experiment was conducted on five consecutive days and comprised a familiarization (A), a training (B), and a test phase (C). **A.** During *familiarization*, listeners heard 132 short German sentences spoken by one male talker without noise-overlay. These sentences consisted of pairs that differed in a single phoneme (either consonant or vowel). On each trial, a total of twelve verbs were presented on a computer screen together with the question *Which word?*, and listeners were instructed to select the verb of the currently presented sentence. **B.** The *training* phase comprised four sessions conducted on the first four days of the experiment. Listeners heard the same sentences spoken by the same talker who was presented during familiarization but this time sentences were presented in noise. On each trial, a total of twenty verbs (five consonant pairs and five vowel pairs) were presented in alphabetic order on a computer screen together with the question *Which word?*, and the current SNR value. Listeners were instructed to select the verb of the currently presented sentence while SRTs were measured using an adaptive tracking procedure. **C.** In the *test* session, listeners were presented with 68 novel sentences spoken by the same talker who was presented in familiarization and training (i.e., familiar talker) and three novel talkers (i.e., unfamiliar talkers). As in training sessions, twenty verbs (five consonant pairs and five vowel pairs) were presented on a computer screen together with the question *Which word?* as well as the current SNR value, and listeners were instructed to select the verb of the currently presented sentence while SRTs were measured using an adaptive tracking procedure.

8.2.3.1 Familiarization Phase

The purpose of the first experimental phase was to familiarize the participants with the voice of one talker (i.e., without noise overlay). The choice of talker was counterbalanced across participants. The familiarization phase was conducted on the first day of testing and comprised a total of 132 sentences (33 consonant pairs and 33 vowel pairs) which were presented in eleven separate blocks. In each block, the verbs of twelve sentences were presented visually on a computer screen together with the question '*Which word?*' (Fig. 8-1A). The corresponding sound files were presented at 65 dB SPL. The order of sounds was randomized. In each trial, participants were asked to move the mouse cursor over the verb that was present in the currently heard sentence and to select this verb by clicking the left mouse button. After each trial, feedback was provided in terms of green (correct) or red (incorrect) coloring of the selected verb. Stimuli that were not correctly identified were repeated at the end of each block. Participants responded in a self-paced manner and could take a short rest after each block. The familiarization procedure lasted about 30 min for each participant.

8.2.3.2 Training Phase

The training comprised four sessions that were performed on four consecutive days. The first training session was performed immediately after the familiarization phase. Each session consisted of twenty blocks. In each block, the verbs of twenty sentences were presented visually on a computer screen together with the question '*Which word?*' and the current SNR value (see below) (Fig. 8-1B). In each block, verbs were selected randomly from the 132 sentences presented during familiarization with the restriction that each block comprised five consonant and five vowel pairs. The verbs were arranged in alphabetic order on a 5 (row) × 4 (column) grid. From these twenty verbs, sound files from the same talker presented during familiarization (i.e., familiar talker) were selected randomly on each trial. Participants were asked to select the verb by moving the mouse cursor over the respective verb and clicking the left mouse button. Sentences were overlaid with speech-shaped noise. The SNR in each block was set initially to +6 dB and was manipulated within a block using a weighted one-up one-down adaptive procedure that estimates SRTs corresponding to 75% correct on the psychometric function (Kaernbach, 1991). In a first phase which lasted until the fifth reversal in the direction of the staircase, SNR was decreased by 2 dB following a correct response and increased by 6 dB following an incorrect response. In a second phase starting with the fifth

reversal, the step sizes were 0.67 dB and 2 dB for down- and up-steps, respectively. Each block ended after the twelfth reversal. SRTs were calculated as the arithmetic mean SNR on the second-phase reversal trials of each block. After each trial, participants received feedback in terms of green (correct) or red (incorrect) coloring of the selected verb. Furthermore, participants could infer their current performance and the difficulty level of the current trial from the SNR value displayed below the verb grid. Participants were instructed to decrease the SNR value as much as possible and that they would gain an additional monetary reward on each day of training if their average SRT for the day was below -5 dB. As in the familiarization phase, participants were allowed to respond at a desired pace and to take a short rest after each block. Each training session lasted about 90 min.

8.2.3.3 Test Phase

The test session was conducted on the fifth day of the study. The experimental procedure was identical to the procedure used in the training phase apart from the following changes: test stimuli comprised 68 novel sentences (17 consonant pairs, 17 vowel pairs), spoken by the same talker as in familiarization and training phase (familiar talker) and three novel talkers (unfamiliar talkers). As in the training sessions, sentences were overlaid with speech-shaped noise and SRTs were measured using the same adaptive tracking procedure. In each block, the verbs of twenty sentences were presented visually on a computer screen together with the question '*Which word?*' and the current SNR value (Fig. 8-1C). Participants were asked to select the verb that was present in the currently heard sentence by moving the mouse cursor over the respective verb and clicking the left mouse button.

There were two different versions of the test session. In the first version, completed by half of the participants, sentences from all four talkers were interleaved in the same block (interleaved paradigm); that is, sound files corresponding to the twenty verbs presented in a given block were selected randomly from all four talkers and four staircases (one per talker) were recorded in each block. Visually presented verbs were selected randomly from the set of test stimuli with the restriction that each block comprised five consonant and five vowel pairs. The interleaved paradigm comprised five blocks. Each block ended when the staircases of all talkers reached twelve reversals. Due to randomization, this could result in more than twelve reversals per talker. However, only the first twelve reversals per talker were analyzed. The other half of participants performed a test session in which sentences from one talker were presented per block (blocked paradigm). The blocked paradigm comprised twenty

blocks. Each talker was presented in five blocks. Identical verb grids, comprising five consonant and five vowel pairs, were used for all talkers to ensure that differences in SRTs between talkers were not due to differences in the presented stimuli. The order of talkers was randomized with the restriction that all four talkers were presented within sequences of four consecutive blocks. Furthermore, it was ensured that there was always a change in talker between two consecutive blocks. There was no restriction on the order of verb grids. In both interleaved and blocked paradigm, feedback was provided immediately after each trial in terms of green (correct) or red (incorrect) coloring of the selected verb. As in the training phase, participants could infer their current performance and the difficulty level of the current trial from the SNR value displayed below the verb grid. In the interleaved paradigm, the mean over current SNR values of all four talkers was presented which limited inferences to the participant's current performance. As in training sessions, participants could gain an additional monetary reward if their average SRT for the day was below -5 dB. Participants were allowed to respond at a desired pace and to take a short rest after each block. The test session lasted about 100 min for each participant.

In summary, the experiment had a 2×2×4 factorial design with the within-subject factor *familiarity* (familiar vs. unfamiliar talker), and the between-subject factors *paradigm* (interleaved vs. blocked) and *familiar talker* (i.e., the talker a given participant was exposed to during training; talker 1–4).

8.3 Results

The results of the training phase are shown in Figure 8-2. A two-way mixed-design analysis of variance (ANOVA) with the within-subject factor training session (session 1–4) and the between-subject factor talker (talker 1–4) on SRTs in the training phase revealed a significant main effect of training session ($F_{(3,54)} = 49.76$; $p < 0.001$) indicating that SRTs decreased over the course of the training (session 1: -6.34 dB, session 2: -6.98 dB, session 3: -7.60 dB, session 4: -8.10 dB). This means that the listeners' abilities to understand speech in noise improved with training. Furthermore, there was a significant main effect of talker ($F_{(3,18)} = 10.90$; $p < 0.001$) indicating that SRTs differed with respect to the talker presented during training (talker 1: -6.49 dB, talker 2 -7.86 dB, talker 3: -8.93 dB, talker 4: -5.82 dB) (Tab. 8-1). This suggests differences in intelligibility across talkers. There was no significant

interaction between training session and familiar talker ($F_{(9,54)} = 0.14$, $p = 0.82$) indicating that listeners' SRTs improved over the course of the training irrespective of which talker they heard.

To compare the thresholds in the last training session with the thresholds at test, we performed a paired samples t-test. This analysis revealed a significant increase in SRTs from the fourth day of training (-8.10 dB) to the test session (-6.93 dB) ($t(23) = -4.05$; $p < 0.001$). Indeed, performance at test almost dropped back to the performance level of the first training session (-6.34 dB). Such a relapse in SRTs is probably due to increased uncertainty induced by the presentation of additional talkers and novel sentences at test.

Table 8-1. Mean SRTs (in dB) with standard deviation in parentheses for the different talkers in training (Day 1 – Day 4) and test (Day 5). SRTs for unfamiliar talkers in test are averaged across all three novel talkers. Familiarity benefit is calculated as the differences between SRTs for familiar vs. unfamiliar talkers in the test session. Note that negative familiarity benefit values denote better comprehension of speech in noise for familiar than unfamiliar talkers.

Talker	Day 1	Day 2	Day 3	Day 4	Day 5 – familiar	Day 5 – unfamiliar	Familiarity Benefit
1	-5.53	-6.07	-6.66	-7.29	-6.54	-6.86	0.32
	(1.56)	(1.21)	(1.51)	(1.73)	(1.66)	(1.41)	(1.04)
2	-6.81	-7.65	-8.30	-8.69	-8.40	-5.82	-2.58
	(0.99)	(1.03)	(1.44)	(0.92)	(1.58)	(1.20)	(0.88)
3	-8.00	-8.72	-9.39	-9.61	-8.38	-7.16	-1.22
	(0.72)	(0.92)	(1.45)	(1.17)	(1.08)	(1.82)	(0.93)
4	-4.60	-5.45	-6.08	-6.80	-5.95	-7.37	1.42
	(0.43)	(0.71)	(0.83)	(0.91)	(1.18)	(0.94)	(1.29)

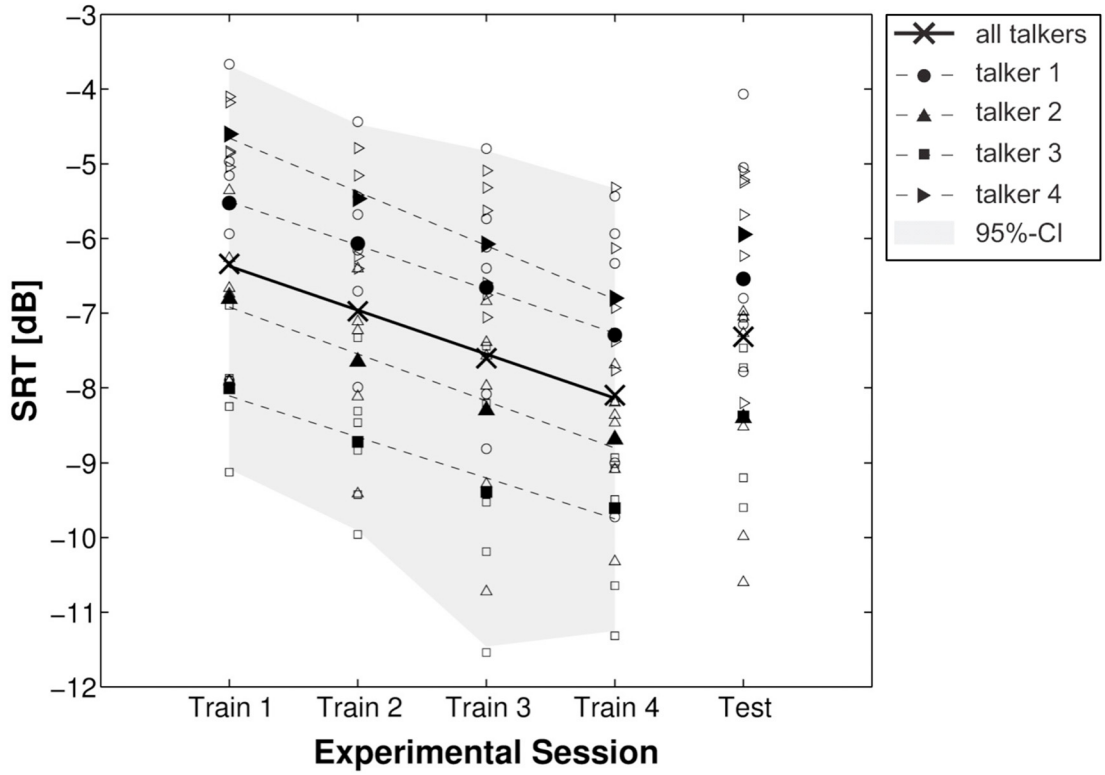


Figure 8-2. Training data. The figure shows mean SRTs in the four training sessions. Black crosses represent mean SRTs averaged across all talkers. Mean SRTs for each talker are coded by different symbols. Linear regression lines are plotted for mean SRTs across all talkers (solid line) and for each talker separately (dashed lines). Individual listeners' SRTs are represented by open symbols. The grey shaded area denotes the 95-% confidence interval. For comparison, the figure also shows mean SRTs in the test session. Note that in the test session all four talkers were presented. Here, symbols only inform about the familiar talker. Each symbol represents the mean SRT averaged across all talkers.

The results of the test session are shown in Figure 8-3. In the test session, listeners heard a set of novel sentences spoken by the same talker who was presented during training (familiar talker) and the three other talkers not presented during training (unfamiliar talkers). Half of the listeners performed the test session in an interleaved manner, where speech from all talkers was presented in each block. The other half performed the test session using a blocked paradigm where speech from only one talker was presented per block. Performance in the test session was analyzed using a three-way mixed-design ANOVA with the within-subject factor talker familiarity (familiar vs. unfamiliar) and the between-subject factors paradigm (interleaved vs. blocked) and familiar talker (talker1–4; i.e., the talker presented during training). The ANOVA revealed a significant main effect of talker familiarity ($F_{(1,16)} = 5.27$, $p = 0.035$) with lower SRTs for familiar (-7.32 dB) than for unfamiliar talkers (-6.80 dB). Hence, on average, listeners were better able to understand speech in noise when it was produced by the familiar talker than by the unfamiliar talkers. The familiarity benefit shown

in Figure 8-3A is the difference in SRTs between familiar and unfamiliar talkers (see below). The main effect of familiar talker was not significant ($F_{(3,16)}$ = 1.02, p = 0.41). This means that SRTs in the test session did not differ with respect to the talker listeners heard during training. There was also no significant main effect of paradigm ($F_{(1,16)}$ = 0.001, p = 0.98) indicating similar SRTs for listeners in the interleaved (-7.07 dB) and blocked paradigm (-7.05 dB) and no significant interaction of paradigm with any of the other factors [paradigm × talker familiarity ($F_{(1,16)}$ = 0.25, p = 0.62); paradigm × familiar talker ($F_{(3,16)}$ = 0.60, p = 0.62); paradigm × talker familiarity × familiar talker ($F_{(3,16)}$ = 0.65, p = 0.59)]. However, there was a significant interaction between talker familiarity and familiar talker ($F_{(3,16)}$ = 15.21; p < 0.001), indicating that the effect of talker familiarity on SRTs depended on which specific talker listeners heard during training.

To investigate the cause of the interaction between talker familiarity and familiar talker, we calculated the familiarity benefit for each listener; that is, the difference in mean SRTs between familiar and unfamiliar talkers (Fig. 8-3A). Note that a *negative* familiarity benefit indicates *better* performance for familiar than for unfamiliar talkers. We then performed one-sample t-tests on the familiarity benefit for listeners in each talker group separately to investigate whether talker-specific familiarity benefits were significantly different from zero (i.e., no familiarity benefit). For listeners who were presented with speech from talker 2 or talker 3 during training, we found a significant familiarity benefit (talker 2: -2.58 dB, t(5) = -7.17, p < 0.001; talker 3: -1.22 dB, t(5) = -3.21, p = 0.024). There was no familiarity benefit for talker 1 (0.32 dB, t(5) = 0.75, p = 0.49). For talker 4, the difference in SRTs between familiar and unfamiliar talkers was significant (1.42, t(5) = 2.68, p = 0.044) but positive; that is, listeners exposed to talker 4 during training showed, on average, higher SRTs for this talker than for the unfamiliar talkers at test. The results of this analysis follow the same pattern as the SRTs in training (i.e., high SRTs for talkers 1 and 4 and low SRTs for talkers 2 and 3) (Fig. 8-2). This suggests that the familiarity benefit depends on the intelligibility of the familiar talker with a benefit for highly intelligible talkers (talkers 2 and 3) and no benefit (talker 1) or even a deficit (talker 4) for less intelligible talkers. Differences in the familiarity benefit across talkers cannot be explained by differential speech recognition performance across listeners in the different talker groups since there was no significant effect of familiar talker on overall SRTs in the test session (see above).

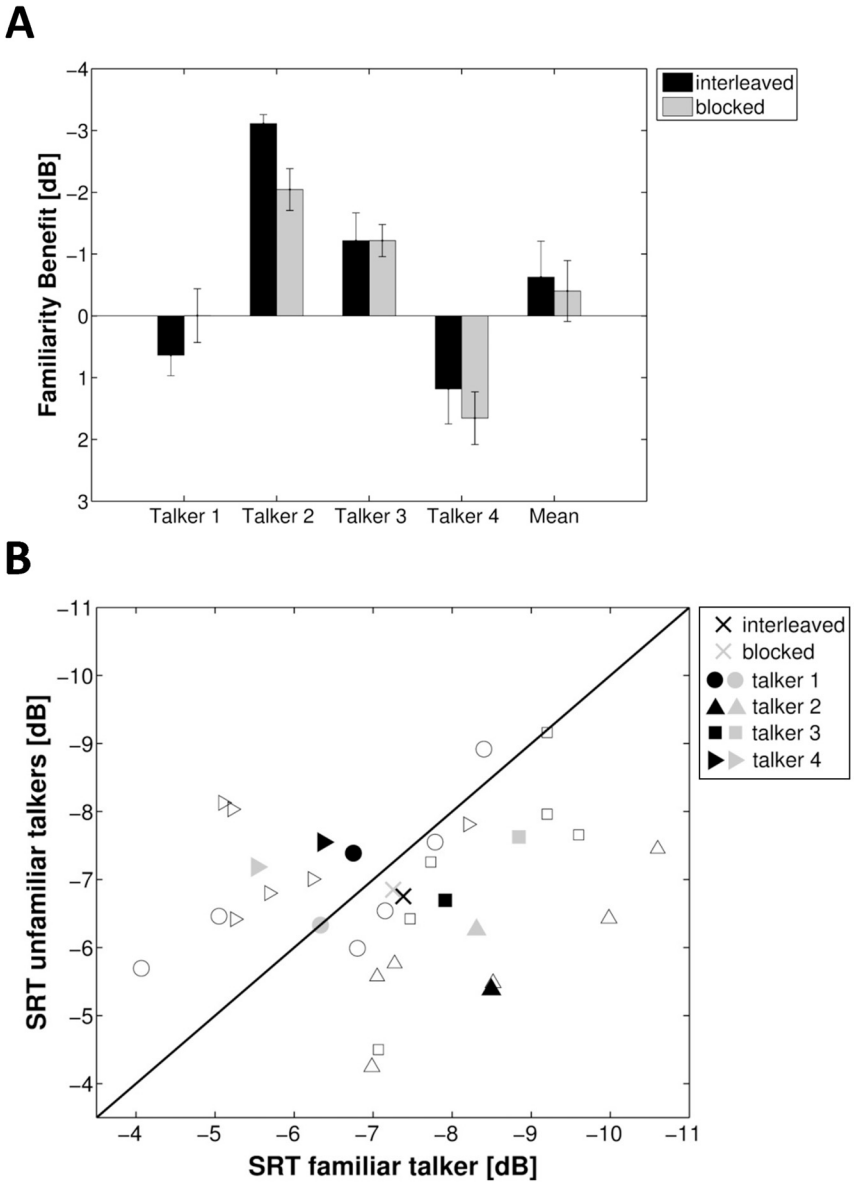


Figure 8-3. Test data. **A.** Familiarity benefit (see section 8.3) is plotted for each talker separately as well as for the mean over all talkers, and separately for interleaved and blocked paradigm. For illustration, the scale of the familiarity benefit is plotted upside down. Negative values denote a familiarity benefit in terms of lower SRTs for familiar than for unfamiliar talkers; positive values denote a familiarity deficit in terms of higher SRTs for familiar than for unfamiliar talkers. Bars show mean familiarity benefit; error bars show 95-% confidence intervals (Morey, 2008). **B.** Mean SRTs for familiar talker are plotted against mean SRTs for unfamiliar talkers. For illustration, the scales of abscissa and ordinate are plotted upside down. Black and grey symbols represent mean SRTs in the interleaved and in the blocked paradigm, respectively. Crosses represent mean SRTs averaged across all familiar talkers. Mean SRTs for each familiar talker are coded by different symbols. Individual listeners' SRTs are represented by open symbols. Symbols below the diagonal indicate that SRTs were lower for familiar than unfamiliar talkers (familiarity benefit). Symbols above the diagonal indicate that SRTs were higher for familiar than unfamiliar talkers (familiarity deficit). The orthogonal distance from the diagonal represents the magnitude of the familiarity effect.

To further investigate differences in talker intelligibility, we checked whether energetic masking was different across talkers by evaluating the "instantaneous" target-to-masker ratio (TMR) on 40 ms chunks of the 200 sentences from each talker mixed with speech-shaped noise at an SNR of 0 dB (Gaudrain and Carlyon, 2013). Although speech sounds were adjusted to the same overall RMS prior to noise masking, it is possible that instantaneous TMR differed across talkers. This is, for example, the case if speech produced by a given talker is more deeply modulated than speech from another talker. A Kruskal-Wallis test revealed significant differences in TMR across talkers ($\chi(3) = 44.68$, $p < 0.001$). Post-hoc Mann-Whitney U-tests showed that TMR was higher for talker 2 and talker 4 than for talker 1 and talker 3 (talker 1 vs. talker 2: $U = 9.13\times10^6$, $p < 0.001$; talker 1 vs. talker 4: $U = 9.69\times10^6$, $p < 0.001$; talker 2 vs. talker 3: $U = 8.29\times10^6$, $p < 0.01$; talker 3 vs. talker 4: $U = 8.93\times10^6$, $p < 0.001$; all other comparisons were not significant) (Suppl. Fig. 8-1; Suppl. Tab. 8-1). Despite the significant results, it is noteworthy that the summary TMR statistics are very similar for all talkers (Suppl. Fig. 8-1E) and that significant TMR differences across talkers are probably due to the large sample size (minimum n = 3,946). Furthermore, differences in TMR across talkers did not occur in the direction that would explain differences in talker intelligibility. In a second analysis, we checked whether talker-specific *f*0 range could account for differences in talker intelligibility (Bradlow et al., 1996). A one-way ANOVA revealed significant differences in *f*0 range across talkers ($F_{(3,766)} = 40.83$, $p < 0.001$). Post-hoc t-tests showed that the *f*0 range is larger for talker 3 than for any of the other talkers (talker 3 vs. 1: $t(382) = 8.47$, $p < 0.001$; talker 3 vs. 2: $t(381) = 9.08$, $p < 0.001$; talker 3 vs. 4: $t(370) = 11.35$, $p < 0.001$; ; all other comparisons were not significant) (Suppl. Fig. 8-2; Suppl. Tab. 8-1). As for TMR, however, talker-specific *f*0 range did not occur in the same direction as differences in talker intelligibility and, therefore, cannot explain differences in the familiarity benefit across talkers.

8.4 Discussion

The results showed that listeners are better able to understand speech in noise when they have prior experience with the talker's voice than when they listen to speech from unfamiliar talkers (Fig. 8-3). However, the occurrence of the familiarity benefit depended on the specific talker the listeners heard during training (Fig. 8-3A). Furthermore, we found that the familiarity benefit can be elicited without explicit voice training. During the first four days of the experiment, listeners were trained to recognize speech in noise from one talker without explicitly directing listeners' attention to the talker dimension. Voice continuity, on the other hand, did not improve speech recognition at test. Listeners who were presented with speech from one talker over a set of sentences showed similar speech recognition performance as listeners for whom talkers randomly changed from trial to trial. Furthermore, voice continuity did not affect the degree to which listeners benefited from talker familiarity (Fig. 8-3B).

A main finding of the present study is that listeners benefit from talker familiarity in speech recognition without explicit voice training. The majority of previous studies investigating talker familiarity used a training procedure in which listeners were explicitly asked to focus on acoustic talker information (Levi et al., 2011; Magnuson et al., 1995; Nygaard et al., 1994; Nygaard and Pisoni, 1998). Here, we employed training that required listeners to focus on the speech message rather than on talker identity. In other words, talker familiarity was induced incidentally while listeners recognized speech in noise. Previous studies provide inconsistent evidence as to whether listeners benefit in speech recognition from such incidental voice training. Yonan and Sommers (2000) found a familiarity benefit for voices that were learned during a task focusing on semantic aspects of sentences. Although their study design did not include genuine voice training, voice recognition scores were assessed after each day of training. This was done by presenting listeners with novel sentences from the same talkers as in the previous training session and from a set of novel talkers. Critically, listeners were asked to classify these sentences as being produced by an "old voice" (i.e., voices presented in the previous training session) or by a "new voice" (i.e., voices not presented during the previous training session). Such task demands required listeners to allocate attention to acoustic voice features. This is an important difference to the training procedure of the present study that did not require listeners to focus on the talker dimension at any time of the experiment. Another study using a training procedure similar to the one employed in the present study failed to find a familiarity benefit following incidental voice training (Burk et

al., 2006). Surprisingly, speech from a familiar talker was *less* likely to be correctly recognized than speech from one of three unfamiliar talkers. Importantly, Burk and colleagues (2006) did not control for differences in talker intelligibility and, unlike the present study, the choice of familiar talker was not counterbalanced across listeners. Furthermore, speech stimuli were presented at a fixed SNR using a method of constant stimuli. Here, using an adaptive tracking procedure, we found that listeners benefited from incidental voice training; that is, listeners showed, on average, lower SRTs for speech that was produced by the same talker who was presented during training than when they listened to novel talkers at test (Fig. 8-3). Our results suggest that listeners implicitly learn acoustic talker information and that they use such information during speech recognition. This is in accordance with previous research showing that talker information is retained and encoded in memory together with the linguistic content of the speech signal (Bradlow et al., 1999; Palmeri et al., 1993; Pisoni, 1993).

Surprisingly, we did not find a stronger familiarity benefit for listeners who performed a test session in which talkers were presented in separate blocks (blocked paradigm) compared to listeners who were presented with speech from randomly changing talkers (interleaved paradigm) (Fig. 8-3B). There is ample evidence that speech recognition is more accurate when the talker is kept constant than when speech from varying talkers is presented (e.g., Bent and Holt, 2013; Best et al., 2008; Bradlow and Pisoni, 1999; Creelman, 1957; Kitterick et al., 2010; Mullennix et al., 1989; Nusbaum and Morin, 1992). Accordingly, one might have expected that listeners in the blocked paradigm would have benefited more from talker familiarity than listeners in the interleaved paradigm. One possible explanation for the finding that there was no influence of blocked talker presentation is that the beneficial effects of voice continuity in the blocked paradigm were canceled out by additional experience with the talkers' voices in the interleaved design. Due to randomization, listeners were presented with more speech from a given talker in the interleaved than in the blocked paradigm (see section 8.2.3.3). Although we only analyzed data from the first twelve reversals, listeners could potentially benefit from additional talker experience in subsequent blocks, thus, compensating for larger talker variability in the interleaved paradigm. Furthermore, the interleaved as well as the blocked paradigm induced talker variability to a certain extent because variations in talkers also occurred in the blocked paradigm albeit on a longer time-scale (i.e., across blocks rather than across trials). Previous studies (e.g., Bent and Holt, 2013; Best et al., 2008; Bradlow and Pisoni, 1999; Nusbaum and Morin, 1992), by contrast,

compared a mixed-talker condition, in which speech from randomly changing talkers was presented, with a single-talker condition without any talker variability at all. In principle, it is also possible that listeners showed a similar familiarity benefit in interleaved and blocked conditions because the blocked presentation of talkers led to lower SRTs for both familiar and unfamiliar talkers. This would, however, imply that SRTs were generally lower in the blocked paradigm. Our results showed that this was not the case (Fig. 8-3B).

In contrast to previous studies that investigated effects of talker familiarity on speech recognition across different groups of listeners (e.g., Nygaard et al., 1994; Nygaard and Pisoni, 1998), this study tested whether speech recognition improves with talker familiarity within a given listener. Testing talker familiarity within listeners inherently controls for any listener-specific effects, such as possible differences in the amount of experience with the experimental procedure or differences in general speech recognition performance across listeners. On the other hand, a within-subject approach comes at the cost of leaving acoustic voice features in the comparison of familiar vs. unfamiliar talkers uncontrolled (Newman and Evers, 2007); that is, effects of talker familiarity in within-subject designs might be due to a variety of vocal characteristics that differ between familiar and unfamiliar talkers. We found that, on average, listeners were better able to understand speech in noise when it was produced by familiar than unfamiliar talkers (Fig. 8-3). Yet, only listeners who were exposed to talker 2 or talker 3 in the training phase benefited from talker familiarity (Fig. 8-3A). Training SRTs were lower for these talkers than for the other two talkers (Fig. 8-2). This suggests that the familiarity benefit depends on the intelligibility of the talker listeners heard during training. Acoustical analyses revealed that differences in talker intelligibility cannot be explained by differences in energetic masking of speech across talkers (Suppl. Fig. 8-1) or talker-specific *f*0 range (Suppl. Fig. 8-2). Talker learning is apparently based on a combination of different acoustic parameters (Lavner et al., 2000) and various global and fine-grained acoustic cues have been shown to play a role in talker intelligibility (Bradlow et al., 1996). Although we used speech from a very homogenous talker group (all male, all native German talkers, all inhabitants of Leipzig at the time of the experiment, and in a narrow age range: 23-31 years), we cannot exclude the possibility that other less quantifiable talker characteristics contributed to differences in talker intelligibility. In general, effects of talker intelligibility are inevitable in within-subject designs but they can be, at least, partially reduced by using familiar and unfamiliar talkers that are roughly equal in intelligibility (Levi et al., 2011; Yonan and Sommers, 2000).

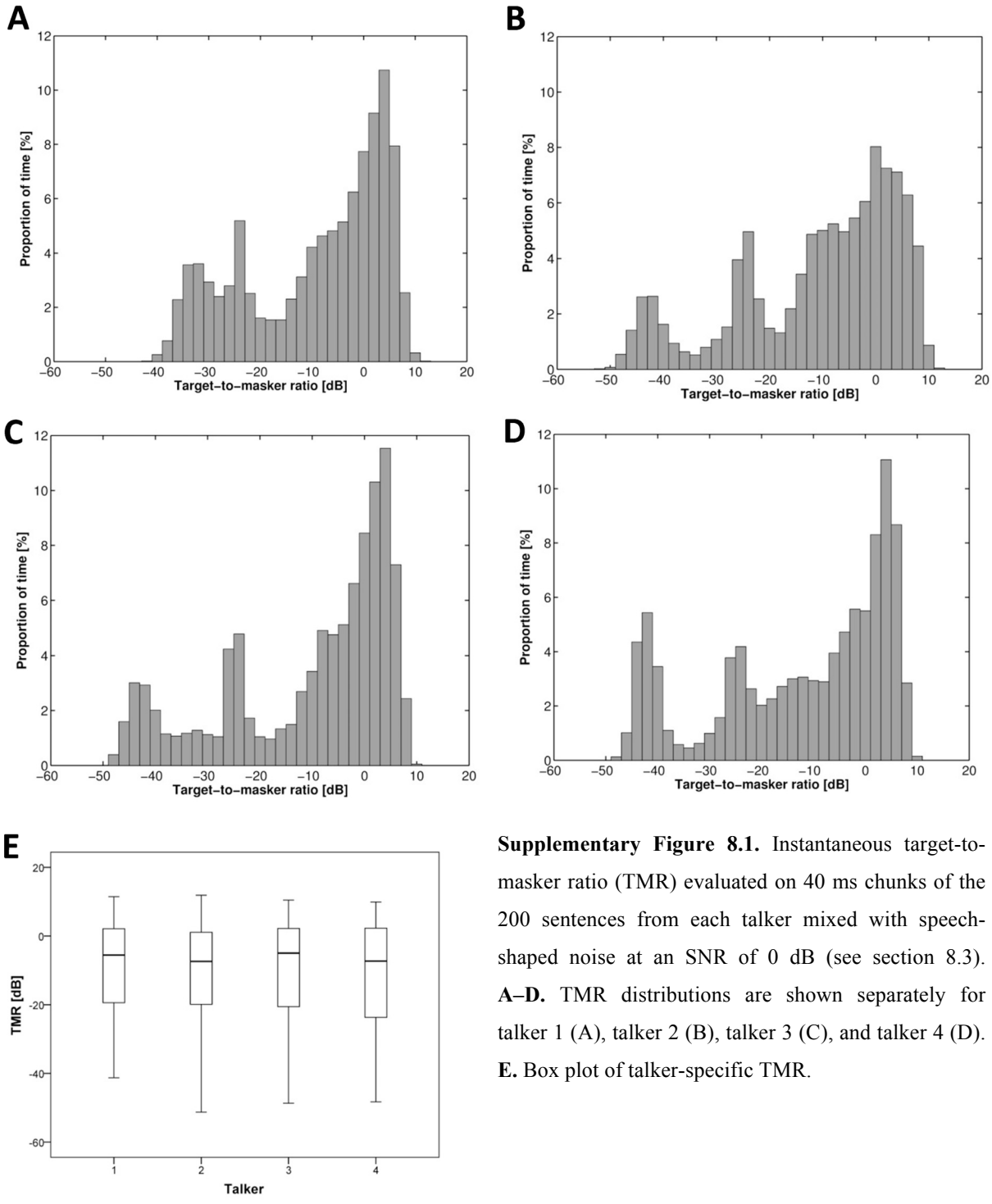
Intuitively, it might appear trivial that we are better able to understand speech when we have prior experience with the talker. This intuition, however, is in stark contrast to the fact that several studies have either failed to show a familiarity benefit in speech comprehension or indicated that mere exposure to a talker's voice is not sufficient to benefit from talker familiarity (Burk et al., 2006; Levi et al., 2011; Magnuson et al., 1995; Nygaard and Pisoni, 1998). Similarly, the results of the present study showed limitations to the familiarity benefit: listeners only benefited from talker familiarity when the talker they heard during training was sufficiently intelligible. The reason for this apparent discrepancy is probably that acquiring knowledge about the vocal characteristics of a talker through natural conversation provides us with more and qualitatively different information than talker learning under laboratory conditions. Indeed, it has been shown that listeners benefit more from personal than experimentally-induced talker familiarity in speech recognition (Magnuson et al., 1995). Recent research suggests that listeners are also better able to understand speech in the presence of competing talkers when it is spoken by their spouses (Johnsrude et al., 2013) and that explicit knowledge about the talker helps stream segregation (Newman and Evers, 2007). In addition to the larger amount of experience listeners usually have with personally familiar talkers than with talkers they are exposed to over a couple of experimental training sessions, natural face-to-face conversation provides the listener with visual talker information which is absent in most experimental settings. Such visual information might, however, help consolidating talker familiarity effects. It has been shown that listeners are better able to transcribe words from auditory speech when they have prior visual experience with the talker during lip-reading (Rosenblum et al., 2007). Furthermore, listeners showed improved recognition of auditory speech when they had previously learned to identify the talker from audio-visual speech samples (i.e., in the presence of both voice and face) (von Kriegstein et al., 2008). Taken together, these findings demonstrate that listeners usually benefit more from personal than experimentally-induced familiarity with a talker in speech recognition and that the larger benefit for personally familiar talkers is possibly based on the larger amount of experience and additional visual information acquired during face-to-face conversation.

Conclusions

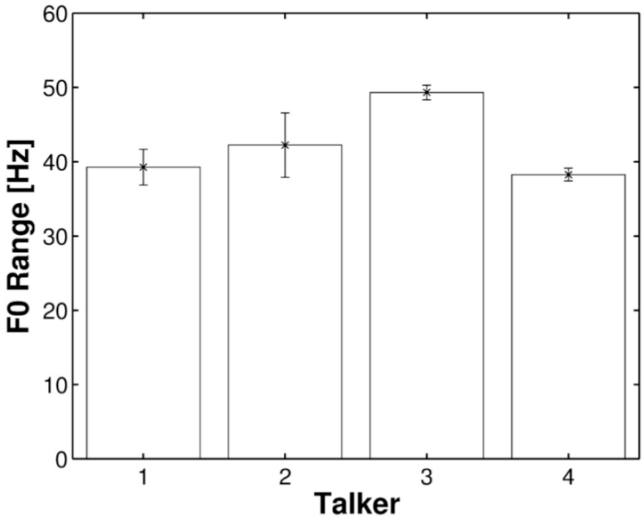
In summary, this study showed that familiarity with a talker's voice enhances linguistic processing. Effects of talker familiarity were demonstrated by lower SRTs for familiar than unfamiliar talkers when recognizing speech in noise. In this study, listeners were familiarized with a talker's voice while performing a linguistic task without being explicitly instructed to focus on talker identity. A benefit in speech recognition from such incidental talker training is consistent with the assumption that linguistic and talker-specific information is processed in an integrated manner (reviewed in Nygaard, 2005) and suggests that listeners implicitly acquire knowledge about the acoustic features specific to a talker's voice during linguistic analysis of the speech signal.

8.5 Supplementary Material

8.5.1 Supplementary Figures



Supplementary Figure 8.1. Instantaneous target-to-masker ratio (TMR) evaluated on 40 ms chunks of the 200 sentences from each talker mixed with speech-shaped noise at an SNR of 0 dB (see section 8.3). **A–D.** TMR distributions are shown separately for talker 1 (A), talker 2 (B), talker 3 (C), and talker 4 (D). **E.** Box plot of talker-specific TMR.



Supplementary Figure 8-2. Bar graph showing talker-specific *f*0 range from 200 sentences (per talker). Bars represent means; error bars show standard error of mean.

8.5.2 Supplementary Tables

Supplementary Table 8-1. Results of acoustical analyses. Geometric mean *f*0 and mean *f*0 range (in Hz) are averaged across 200 sentences from each talker with standard deviation in parentheses. Target-to-masker ratio (TMR) was evaluated on 40 ms chunks of 200 target sentences for each talker (see section 8.3). Talker-specific median TMRs are shown with 95-% confidence intervals in parentheses.

Talker	*f*0 mean [Hz]	*f*0 range [Hz]	TMR [dB]
1	85.25	36.67	-5.55
	(4.72)	(13.65)	(42.70)
2	79.63	36.25	-7.39
	(4.32)	(13.00)	(52.50)
3	124.26	47.12	-4.95
	(4.48)	(10.12)	(51.49)
4	91.00	35.96	-7.30
	(2.74)	(8.73)	(51.31)

References

Abberton, E., Fourcin, A.J., 1978. Intonation and speaker identification. Language and Speech 21, 305-318.

Abramson, A.S., Cooper, F.S., 1959. Perception of American English vowels in terms of a reference system. Haskins Laboratories Quarterly Progress Report QPR-32, Appendix, 1.

Ainsworth, W.A., 1975. Intrinsic and extrinsic factors in vowel judgments. Auditory analysis and perception of speech, 103-113.

Andics, A., McQueen, J.M., Petersson, K.M., Gal, V., Rudas, G., Vidnyanszky, Z., 2010. Neural mechanisms for voice recognition. Neuroimage 52, 1528-1540.

Assal, G., Aubert, C., Buttet, J., 1981. Asymétrie cérébrale et reconnaissance de la voix. Revue Neurologique 137, 255-268.

Assal, G., Zander, E., Kremin, H., Buttet, J., 1976. Discrimination des voix lors des lesions du cortex cerebral. Schweiz. Arch. Neurol. Neurochir. Psychiatr 119, 307-315.

Baumann, O., Belin, P., 2010. Perceptual scaling of voice identity: common dimensions for different vowels and speakers. Psychological Research PRPF 74, 110-120.

Belin, P., Fecteau, S., Bedard, C., 2004. Thinking the voice: neural correlates of voice perception. Trends in Cognitive Sciences 8, 129-135.

Belin, P., Zatorre, R.J., 2003. Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport 14, 2105-2109.

Belin, P., Zatorre, R.J., Ahad, P., 2002. Human temporal-lobe response to vocal sounds. Cognitive Brain Research 13, 17-26.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309-312.

Bent, T., Holt, R.F., 2013. The influence of talker and foreign-accent variability on spoken word identification. Journal of the Acoustical Society of America 133, 1677-1686.

Best, V., Ozmeral, E.J., Kopco, N., Shinn-Cunningham, B.G., 2008. Object continuity enhances selective auditory attention. Proc Natl Acad Sci U S A 105, 13174-13178.

Bradlow, A.R., Nygaard, L.C., Pisoni, D.B., 1999. Effects of talker, rate, and amplitude variation on recognition memory for spoken words. Perception & Psychophysics 61, 206-219.

Bradlow, A.R., Pisoni, D.B., 1999. Recognition of spoken words by native and non-native listeners: talker-, listener-, and item-related factors. Journal of the Acoustical Society of America 106, 2074-2085.

Bradlow, A.R., Torretta, G.M., Pisoni, D.B., 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. Speech Communication 20, 255-272.

Burk, M.H., Humes, L.E., Amos, N.E., Strauser, L.E., 2006. Effect of training on word-recognition performance in noise for young normal-hearing and older hearing-impaired listeners. Ear and hearing 27, 263-278.

Chandrasekaran, B., Chan, A.H.D., Wong, P.C.M., 2011. Neural Processing of What and Who Information in Speech. Journal of Cognitive Neuroscience 23, 2690-2700.

Clopper, C.G., Pisoni, D.B., De Jong, K., 2005. Acoustic characteristics of the vowel systems of six regional varieties of American English. Journal of the Acoustical Society of America 118, 1661-1676.

Couper-Kuhlen, E., 1986. An introduction to English prosody. Edward Arnold.

Creelman, C.D., 1957. Case of the unknown talker. Journal of the Acoustical Society of America 29, 655-655.

Fant, G., 1960. Acoustic Theory of Speech Production. Mouton & Co, The Hague, Netherlands.

Fecteau, S., Armony, J.L., Joanette, Y., Belin, P., 2004. Is voice processing species-specific in human auditory cortex? An fMRI study. Neuroimage 23, 840-848.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. " Who" Is Saying" What"? Brain-Based Decoding of Human Voice and Speech. Science 322, 970-973.

Frick, R.W., 1985. Communicating emotion: The role of prosodic features. Psychological Bulletin 97, 412-429.

Friederici, A.D., 2002. Towards a neural basis of auditory sentence processing. Trends in Cognitive Sciences 6, 78-84.

Friederici, A.D., 2011. The Brain Basis of Language Processing: From Structure to Function. Physiological Reviews 91, 1357-1392.

Friederici, A.D., Alter, K., 2004. Lateralization of auditory language functions: A dynamic dual pathway model. Brain and Language 89, 267-276.

Friston, K.J., Buechel, C., Fink, G.R., Morris, J., Rolls, E., Dolan, R.J., 1997. Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6, 218-229.

Gandour, J., Wong, D., Hsieh, L., Weinzapfel, B., Van Lancker, D., Hutchins, G.D., 2000. A crosslinguistic PET study of tone perception. J Cogn Neurosci 12, 207-222.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J.R., Schweinberger, S.R., Warren, J.D., Duchaine, B., 2009. Developmental phonagnosia: a selective deficit of vocal identity recognition. Neuropsychologia 47, 123-131.

Gaudrain, E., Carlyon, R.P., 2013. Using Zebra-speech to study sequential and simultaneous speech segregation in a cochlear-implant simulation. Journal of the Acoustical Society of America, 133, 502-518.

Gaudrain, E., Li, S., Ban, V.S., Patterson, R.D., 2009. The role of glottal pulse rate and vocal tract length in the perception of speaker identity. Interspeech 2009: 10th Annual Conference of the International Speech Communication Association 2009, Vols 1-5, 152-155.

Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Sfaello, I., Barthélémy, C., Brunelle, F., Samson, Y., Zilbovicius, M., 2004. Abnormal cortical voice processing in autism. Nature neuroscience 7, 801-802.

Goldinger, S.D., 1998. Echoes of echoes? An episodic theory of lexical access. Psychological review 105, 251-279.

Goldinger, S.D., Pisoni, D.B., Logan, J.S., 1991. On the nature of talker variability effects on recall of spoken word lists. Journal of Experimental Psychology: Learning, Memory, and Cognition 17, 152-162.

Griffiths, T.D., Kumar, S., Sedley, W., Nourski, K.V., Kawasaki, H., Oya, H., Patterson, R.D., Brugge, J.F., Howard, M.A., 2010. Direct Recordings of Pitch Responses from Human Auditory Cortex. Current Biology 20, 1128-1132.

Hagiwara, R., 1997. Dialect variation and formant frequency: The American English vowels revisited. Journal of the Acoustical Society of America 102, 655-658.

Hickok, G., Poeppel, D., 2007. Opinion - The cortical organization of speech processing. Nature Reviews Neuroscience 8, 393-402.

Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K., 1995. Acoustic characteristics of American English vowels. Journal of the Acoustical society of America 97, 3099-3111.

Howie, J.M., 1976. Acoustical studies of Mandarin vowels and tones. Cambridge University Press.

Johnsrude, I.S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H.P., Carlyon, R.P., 2013. Swinging at a Cocktail Party Voice Familiarity Aids Speech Perception in the Presence of a Competing Voice. Psychological science 24, 1995-2004.

Joos, M. (1948). Acoustic phonetics. Language 24, 5-136.

Kaernbach, C., 1991. Simple Adaptive Testing with the Weighted up-down Method. Percept Psychophys 49, 227-229.

Kaganovich, N., Francis, A.L., Melara, R.D., 2006. Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. Brain Research 1114, 161-172.

Kawahara, H., Irino, T., 2004. Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. Speech separation by humans and machines, 167-180.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H., 2008. Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Vols 1-12, 3933-3936.

Kitterick, P.T., Bailey, P.J., Summerfield, A.Q., 2010. Benefits of knowing who, where, and when in multi-talker listening. Journal of the Acoustical Society of America 127, 2498-2508.

Kreiman, J., Van Lancker-Sidtis, D., Gerratt, B.R., 2005. Perception of voice quality. The handbook of speech perception, 338-362.

Kumar, S., Sedley, W., Nourski, K.V., Kawasaki, H., Oya, H., Patterson, R.D., Howard, M.A., Friston, K.J., Griffiths, T.D., 2011. Predictive Coding and Pitch Processing in the Auditory Cortex. Journal of Cognitive Neuroscience 23, 3084-3094.

Labov, W., 1972. Sociolinguistic patterns. University of Pennsylvania Press.

Ladefoged, P., Broadbent, D.E., 1957. Information conveyed by vowels. Journal of the Acoustical Society of America 29, 98-104.

Lass, N.J., Hughes, K.R., Bowyer, M.D., Waters, L.T., Bourne, V.T., 1976. Speaker sex identification from voiced, whispered, and filtered isolated vowels. Journal of the Acoustical Society of America 59, 675-678.

Latinus, M., Crabbe, F., Belin, P., 2011. Learning-Induced Changes in the Cerebral Processing of Voice Identity. Cerebral Cortex 21, 2820-2828.

Lavner, Y., Gath, I., Rosenhouse, J., 2000. The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. Speech Communication 30, 9-26.

Leff, A.P., Schofield, T.M., Stephan, K.E., Crinion, J.T., Friston, K.J., Price, C.J., 2008. The Cortical Dynamics of Intelligible Speech. Journal of Neuroscience 28, 13209-13215.

Levi, S.V., Winters, S.J., Pisoni, D.B., 2011. Effects of cross-language voice training on speech perception: whose familiar voices are more intelligible? Journal of the Acoustical Society of America 130, 4053-4062.

Magnuson, J.S., Nusbaum, H.C., 2007. Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. Journal of Experimental Psychology-Human Perception and Performance 33, 391-409.

Magnuson, J.S., Yamada, R.A., Nusbaum, H.C., 1995. The effects of familiarity with a voice on speech perception. In Proceedings of the 1995 Spring Meeting of the Acoustical Society of Japan (pp. 391-392).

Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., Corthals, P., 2009. Acoustic measurement of overall voice quality: A meta-analysis. Journal of the Acoustical Society of America 126, 2619-2634.

Matsumoto, H., Hiki, S, Sone, T., Nimura, T., 1973. Multidimensional representation of personal quality of vowels and its acoustical correlates. IEEE Transactions on Audio and Electroacoustics 21, 428-436.

McGettigan, C., Scott, S.K., 2012. Cortical asymmetries in speech perception: what's wrong, what's right and what's left? Trends in Cognitive Sciences 16, 269-276.

Meyer, M., Alter, K., Friederici, A.D., Lohmann, G., von Cramon, D.Y., 2002. FMRI reveals brain regions mediating slow prosodic modulations in spoken sentences. Human Brain Mapping 17, 73-88.

Morey, R.D., 2008. Confidence intervals from normalized data: A correction to Cousineau (2005). Tutorials in Quantitative Methods for Psychology 4, 61-64.

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., Zilles, K., 2001. Human primary auditory cortex: Cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage 13, 684-701.

Mullennix, J.W., Pisoni, D.B., Martin, C.S., 1989. Some effects of talker variability on spoken word recognition. Journal of the Acoustical Society of America 85, 365-378.

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., Nagumo, S., Kubota, K., Fukuda, H., Ito, K., Kojima, S., 2001. Neural substrates for recognition of familiar voices: a PET study. Neuropsychologia 39, 1047-1054.

Nearey, T.M., 1989. Static, dynamic, and relational properties in vowel perception. Journal of the Acoustical Society of America 85, 2088- 2113.

Newman, R.S., Evers, S., 2007. The effect of talker familiarity on stream segregation. Journal of Phonetics 35, 85-103.

Nusbaum, H., Magnuson, J., 1997. Talker normalization: Phonetic constancy as a cognitive process. Talker variability in speech processing, 109-132.

Nusbaum, H.C., Morin, T.M., 1992. Paying attention to differences among talkers. Speech perception, production and linguistic structure, 113-134.

Nygaard, L.C., 2005. Perceptual integration of linguistic and nonlinguistic properties of speech. The handbook of speech perception, 390-413.

Nygaard, L.C., Pisoni, D.B., 1998. Talker-specific learning in speech perception. Perception & Psychophysics 60, 355-376.

Nygaard, L.C., Sommers, M.S., Pisoni, D.B., 1994. Speech-Perception as a Talker-Contingent Process. Psychological Science 5, 42-46.

O'Shaughnessy, D., 2008. Invited paper: Automatic speech recognition: History, methods and challenges. Pattern Recognition 41, 2965-2979.

Oldfield, R.C., 1971. The Assessment and Analysis of Handedness: The Edinburgh Inventory. Neuropsychologia 9, 97-113.

Palmeri, T.J., Goldinger, S.D., Pisoni, D.B., 1993. Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words. Journal of Experimental Psychology-Learning Memory and Cognition 19, 309-328.

Perrachione, T.K., Del Tufo, S.N., & Gabrieli, J.D., 2011. Human voice recognition depends on language ability. Science 333, 595-595.

Peterson, G.E., Barney, H.L., 1952. Control Methods Used in a Study of the Vowels. Journal of the Acoustical Society of America 24, 175-184.

Pisoni, D.B., 1993. Long-Term-Memory in Speech-Perception - Some New Findings on Talker Variability, Speaking Rate and Perceptual-Learning. Speech Communication 13, 109-125.

Pisoni, D.B., 1997. Some thoughts on “normalization” in speech perception. Talker variability in speech processing, 9-32.

Plante, E., Creusere, M., Sabin, C., 2002. Dissociating sentential prosody from sentence processing: activation interacts with task demands. Neuroimage 17, 401-410.

Remez, R.E., Fellowes, J.M., Rubin, P.E., 1997. Talker identification based on phonetic information. Journal of Experimental Psychology: Human Perception and Performance 23, 651-666.

Rosenblum, L.D., Miller, R.M., Sanchez, K., 2007. Lip-Read Me Now, Hear Me Better Later Cross-Modal Transfer of Talker-Familiarity Effects. Psychological Science 18, 392-396.

Salvata, C., Blumstein, S.E., Myers, E.B., 2012. Speaker invariance for phonetic information: An fMRI investigation. Language and cognitive processes 27, 210-230.

Schweinberger, S.R., Walther, C., Zaske, R., Kovacs, G., 2011. Neural correlates of adaptation to voice identity. British Journal of Psychology 102, 748-764.

Scott, S.K., 2005. Auditory processing - speech, space and auditory objects. Current Opinion in Neurobiology 15, 197-201.

Scott, S.K., Johnsrude, I.S., 2003. The neuroanatomical and functional organization of speech perception. Trends in Neurosciences 26, 100-107.

Shankweiler, D.P., Strange, W., Verbrugge, R.R., 1977. Speech and the problem of perceptual constancy. Perceiving, acting, and knowing: Toward an ecological psychology, 315-345.

Shipp, T., Hollien, H., 1969. Perception of the aging male voice. Journal of Speech & Hearing Research 12, 703-710.

Stevens, K.N., House, A.S., 1961. An acoustical theory of vowel production and some of its implications. Journal of Speech, Language and Hearing Research 4, 303-320.

Syrdal, A.K., Gopal, H.S., 1986. A perceptual model of vowel recognition based on the auditory representation of American English vowels. Journal of the Acoustical Society of America 79, 1086-1100.

Van Lancker, D.R., 1980. Cerebral lateralization of pitch cues in the linguistic signal. Research on Language & Social Interaction 13, 201-277.

Van Lancker, D.R., Canter, G.J., 1982. Impairment of voice and face recognition in patients with hemispheric damage. Brain and Cognition 1, 185-195.

Van Lancker, D.R., Kreiman, J., Cummings, J., 1989. Voice perception deficits: Neuroanatomical correlates of phonagnosia. Journal of Clinical and Experimental Neuropsychology 11, 665-674.

Van Lancker, D., Kreiman, J., Emmorey, K., 1985. Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. Journal of Phonetics 13, 19-38.

Vigneau, M., Beaucousin, V., Herve, P.Y., Duffau, H., Crivello, F., Houde, O., Mazoyer, B., Tzourio-Mazoyer, N., 2006. Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. Neuroimage 30, 1414-1432.

von Kriegstein, K., Dogan, Ö., Grüter, M., Giraud, A.L., Kell, C. A., Grüter, T., Kleinschmidt, A., Kiebel, S.J., 2008. Simulation of talking faces in the human brain improves auditory speech recognition. Proceedings of the National Academy of Sciences 105, 6747-6752.

von Kriegstein, K., Eger, E., Kleinschmidt, A., Giraud, A.L., 2003. Modulation of neural responses to speech by directing attention to voices or verbal content. Cognitive Brain Research 17, 48-55.

von Kriegstein, K., Giraud, A.L., 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage 22, 948-955.

von Kriegstein, K., Smith, D.R.R., Patterson, R.D., Ives, D.T., Griffiths, T.D., 2007. Neural representation of auditory size in the human voice and in sounds from other resonant sources. Current Biology 17, 1123-1128.

von Kriegstein, K., Smith, D.R.R., Patterson, R.D., Kiebel, S.J., Griffiths, T.D., 2010. How the Human Brain Recognizes Speech in the Context of Changing Speakers. Journal of Neuroscience 30, 629-638.

von Kriegstein, K., Warren, J.D., Ives, D.T., Patterson, R.D., Griffiths, T.D., 2006. Processing the acoustic effect of size in speech sounds. Neuroimage 32, 368-375.

Walden, B.E., Montgomery, A.A., Gibeily, G.J., Prosek, R.A., Schwartz, D.M., 1978. Correlates of psychological dimensions in talker similarity. Journal of Speech, Language and Hearing Research 21, 265-275.

Wildgruber, D., Hertrich, I., Riecker, A., Erb, M., Anders, S., Grodd, W., Ackermann, H., 2004. Distinct frontal regions subserve evaluation of linguistic and emotional aspects of speech intonation. Cerebral Cortex 14, 1384-1389.

Witteman, J., van Ijzendoorn, M.H., van de Velde, D., van Heuven, V.J.J.P., Schiller, N.O., 2011. The nature of hemispheric specialization for linguistic and emotional prosodic perception: A meta-analysis of the lesion literature. Neuropsychologia 49, 3722-3738.

Wong, P.C.M., Nusbaum, H.C., Small, S.L., 2004. Neural bases of talker normalization. Journal of Cognitive Neuroscience 16, 1173-1184.

Wong, P.C.M., Warrier, C.M., Penhune, V.B., Roy, A.K., Sadehh, A., Parrish, T.B., Zatorre, R.J., 2008. Volume of left heschl's gyrus and linguistic pitch learning. Cerebral Cortex 18, 828-836.

Yonan, C.A., Sommers, M.S., 2000. The effects of talker familiarity on spoken word identification in younger and older listeners. Psychology and Aging 15, 88-99.

Appendix

A. List of Figures

B. List of Supplementary Figures

C. List of Tables

D. List of Supplementary Tables

Versicherung über die selbständige Erarbeitung der Dissertation

Hiermit erkläre ich, dass die vorliegende Arbeit ohne unzulässige Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt wurde und dass die aus fremden Quellen direkt oder indirekt übernommenen Gedanken in der Arbeit als solche kenntlich gemacht worden sind.

Jens Kreitewolf

Leipzig, den 10.12.2013